PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, G06F 15/00 — A1

(11) International Publication Number: **WO 95/20681**

(43) International Publication Date: 3 August 1995 (03.08.95)

(21) International Application Number: PCT/US95/01160

(22) International Filing Date: 27 January 1995 (27.01.95)

(30) Priority Data:

08/187,530	27 January 1994 (27.01.94)	US
08/282,955	29 July 1994 (29.07.94)	US

(71) Applicant: INCYTE PHARMACEUTICALS, INC. [US/US]; 3330 Hillview Avenue, Palo Alto, CA 94304 (US).

(72) Inventors: SEILHAMER, Jeffrey, J.; 12555 La Cresta, Los Altos Hills, CA 94022 (US). SCOTT, Randal, W.; 13140 Sun-Mor, Mountain View, CA 94040 (US).

(74) Agents: CAGE, Kenneth, L. et al.; Willian Brinks Hofer Gilson & Lione, 2000 K Street, N.W., Suite 200, Washington, DC 20006-1809 (US).

(81) Designated States: AM, AU, BB, BG, BR, BY, CA, CN, CZ, EE, FI, GE, HU, JP, KG, KP, KR, KZ, LK, LR, LT, LV, MD, MG, MN, MX, NO, NZ, PL, RO, RU, SI, SK, TJ, TT, UA, UZ, VN, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG), ARIPO patent (KE, MW, SD, SZ).

Published
With international search report.

(54) Title: COMPARATIVE GENE TRANSCRIPT ANALYSIS

(57) Abstract

A method and system for quantifying the relative abundance of gene transcripts in a biological specimen. One embodiment of the method generates high-throughput sequence-specific analysis of multiple RNAs or their corresponding cDNAs (gene transcript imaging analysis). Another embodiment of the method produces a gene transcript imaging analysis by the use of high-throughput cDNA sequence analysis. In addition, the gene transcript imaging can be used to detect or diagnose a particular biological state, disease, or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells. The invention provides a method for comparing the gene transcript image analysis from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes which are differentially expressed between the two specimens.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	**GB**	United Kingdom	**MR**	Mauritania
AU	Australia	**GE**	Georgia	**MW**	Malawi
BB	Barbados	**GN**	Guinea	**NE**	Niger
BE	Belgium	**GR**	Greece	**NL**	Netherlands
BF	Burkina Faso	**HU**	Hungary	**NO**	Norway
BG	Bulgaria	**IE**	Ireland	**NZ**	New Zealand
BJ	Benin	**IT**	Italy	**PL**	Poland
BR	Brazil	**JP**	Japan	**PT**	Portugal
BY	Belarus	**KE**	Kenya	**RO**	Romania
CA	Canada	**KG**	Kyrgystan	**RU**	Russian Federation
CF	Central African Republic	**KP**	Democratic People's Republic of Korea	**SD**	Sudan
CG	Congo			**SE**	Sweden
CH	Switzerland	**KR**	Republic of Korea	**SI**	Slovenia
CI	Côte d'Ivoire	**KZ**	Kazakhstan	**SK**	Slovakia
CM	Cameroon	**LI**	Liechtenstein	**SN**	Senegal
CN	China	**LK**	Sri Lanka	**TD**	Chad
CS	Czechoslovakia	**LU**	Luxembourg	**TG**	Togo
CZ	Czech Republic	**LV**	Latvia	**TJ**	Tajikistan
DE	Germany	**MC**	Monaco	**TT**	Trinidad and Tobago
DK	Denmark	**MD**	Republic of Moldova	**UA**	Ukraine
ES	Spain	**MG**	Madagascar	**US**	United States of America
FI	Finland	**ML**	Mali	**UZ**	Uzbekistan
FR	France	**MN**	Mongolia	**VN**	Viet Nam
GA	Gabon				

COMPARATIVE GENE TRANSCRIPT ANALYSIS

1. FIELD OF INVENTION

The present invention is in the field of molecular biology and computer science; more particularly, the present invention describes methods of analyzing gene transcripts and diagnosing the genetic expression of cells and tissue.

2. BACKGROUND OF THE INVENTION

Until very recently, the history of molecular biology has been written one gene at a time. Scientists have observed the cell's physical changes, isolated mixtures from the cell or its milieu, purified proteins, sequenced proteins and therefrom constructed probes to look for the corresponding gene.

Recently, different nations have set up massive projects to sequence the billions of bases in the human genome. These projects typically begin with dividing the genome into large portions of chromosomes and then determining the sequences of these pieces, which are then analyzed for identity with known proteins or portions thereof, known as motifs. Unfortunately, the majority of genomic DNA does not encode proteins and though it is postulated to have some effect on the cell's ability to make protein, its relevance to medical applications is not understood at this time.

A third methodology involves sequencing only the transcripts encoding the cellular machinery actively involved in making protein, namely the mRNA. The advantage is that the cell has already edited out all the non-coding DNA, and it is relatively easy to identify the protein-coding portion of the RNA. The utility of this approach was not immediately obvious to genomic researchers. In fact, when cDNA sequencing was initially proposed, the method was roundly denounced by those committed to genomic sequencing. For example, the head of the U.S. Human Genome project discounted CDNA sequencing as not valuable and refused to approve funding of projects.

In this disclosure, we teach methods for analyzing DNA, including cDNA libraries. Based on our analyses and

research, we see each individual gene product as a "pixel" of information, which relates to the expression of that, and only that, gene. We teach herein, methods whereby the individual "pixels" of gene expression information can be combined into a single gene transcript "image," in which each of the individual genes can be visualized simultaneously and allowing relationships between the gene pixels to be easily visualized and understood.

We further teach a new method which we call electronic subtraction. Electronic subtraction will enable the gene researcher to turn a single image into a moving picture, one which describes the temporality or dynamics of gene expression, at the level of a cell or a whole tissue. It is that sense of "motion" of cellular machinery on the scale of a cell or organ which constitutes the new invention herein. This constitutes a new view into the process of living cell physiology and one which holds great promise to unveil and discover new therapeutic and diagnostic approaches in medicine.

We teach another method which we call "electronic northern," which tracks the expression of a single gene across many types of cells and tissues.

Nucleic acids (DNA and RNA) carry within their sequence the hereditary information and are therefore the prime molecules of life. Nucleic acids are found in all living organisms including bacteria, fungi, viruses, plants and animals. It is of interest to determine the relative abundance of different discrete nucleic acids in different cells, tissues and organisms over time under various conditions, treatments and regimes.

All dividing cells in the human body contain the same set of 23 pairs of chromosomes. It is estimated that these autosomal and sex chromosomes encode approximately 100,000 genes. The differences among different types of cells are believed to reflect the differential expression of the 100,000 or so genes. Fundamental questions of biology could be answered by understanding which genes are transcribed and knowing the relative abundance of transcripts in different cells.

Previously, the art has only provided for the analysis of a few known genes at a time by standard molecular biology techniques such as PCR, northern blot analysis, or other types of DNA probe analysis such as in situ hybridization. Each of these methods allows one to analyze the transcription of only known genes and/or small numbers of genes at a time. Nucl. Acids Res. 19, 7097-7104 (1991); Nucl. Acids Res. 18, 4833-42 (1990); Nucl. Acids Res. 18, 2789-92 (1989); European J. Neuroscience 2, 1063-1073 (1990); Analytical Biochem. 187, 364-73 (1990); Genet. Annals Techn. Appl. 7, 64-70 (1990); GATA 8(4), 129-33 (1991); Proc. Natl. Acad. Sci. USA 85, 1696-1700 (1988); Nucl. Acids Res. 19, 1954 (1991); Proc. Natl. Acad. Sci. USA 88, 1943-47 (1991); Nucl. Acids Res. 19, 6123-27 (1991); Proc. Natl. Acad. Sci. USA 85, 5738-42 (1988); Nucl. Acids Res. 16, 10937 (1988).

Studies of the number and types of genes whose transcription is induced or otherwise regulated during cell processes such as activation, differentiation, aging, viral transformation, morphogenesis, and mitosis have been pursued for many years, using a variety of methodologies. One of the earliest methods was to isolate and analyze levels of the proteins in a cell, tissue, organ system, or even organisms both before and after the process of interest. One method of analyzing multiple proteins in a sample is using 2-dimensional gel electrophoresis, wherein proteins can be, in principle, identified and quantified as individual bands, and ultimately reduced to a discrete signal. At present, 2-dimensional analysis only resolves approximately 15% of the proteins. In order to positively analyze those bands which are resolved, each band must be excised from the membrane and subjected to protein sequence analysis using Edman degradation. Unfortunately, most of the bands were present in quantities too small to obtain a reliable sequence, and many of those bands contained more than one discrete protein. An additional difficulty is that many of the proteins were blocked at the amino-terminus, further complicating the sequencing process.

Analyzing differentiation at the gene transcription level has overcome many of these disadvantages and drawbacks, since the power of recombinant DNA technology allows amplification of signals containing very small amounts of material. The most common method, called "hybridization subtraction," involves isolation of mRNA from the biological specimen before (B) and after (A) the developmental process of interest, transcribing one set of mRNA into cDNA, subtracting specimen B from specimen A (mRNA from cDNA) by hybridization, and constructing a cDNA library from the non-hybridizing mRNA fraction. Many different groups have used this strategy successfully, and a variety of procedures have been published and improved upon using this same basic scheme. Nucl. Acids Res. 19, 7097-7104 (1991); Nucl. Acids Res. 18, 4833-42 (1990); Nucl. Acids Res. 18, 2789-92 (1989); European J. Neuroscience 2, 1063-1073 (1990); Analytical Biochem. 187, 364-73 (1990); Genet. Annals Techn. Appl. 7, 64-70 (1990); GATA 8(4), 129-33 (1991); Proc. Natl. Acad. Sci. USA 85, 1696-1700 (1988); Nucl. Acids Res. 19, 1954 (1991); Proc. Natl. Acad. Sci. USA 88, 1943-47 (1991); Nucl. Acids Res. 19, 6123-27 (1991); Proc. Natl. Acad. Sci. USA 85, 5738-42 (1988); Nucl. Acids Res. 16, 10937 (1988).

Although each of these techniques have particular strengths and weaknesses, there are still some limitations and undesirable aspects of these methods: First, the time and effort required to construct such libraries is quite large. Typically, a trained molecular biologist might expect construction and characterization of such a library to require 3 to 6 months, depending on the level of skill, experience, and luck. Second, the resulting subtraction libraries are typically inferior to the libraries constructed by standard methodology. A typical conventional cDNA library should have a clone complexity of at least 10^6 clones, and an average insert size of 1-3 kB. In contrast, subtracted libraries can have complexities of 10^2 or 10^3 and average insert sizes of 0.2 kB. Therefore, there can be a significant loss of clone and sequence information associated with such libraries. Third, this

approach allows the researcher to capture only the genes induced in specimen A relative to specimen B, not vice-versa, nor does it easily allow comparison to a third specimen of interest (C). Fourth, this approach requires very large amounts (hundreds of micrograms) of "driver" mRNA (specimen B), which significantly limits the number and type of subtractions that are possible since many tissues and cells are very difficult to obtain in large quantities.

Fifth, the resolution of the subtraction is dependent upon the physical properties of DNA:DNA or RNA:DNA hybridization. The ability of a given sequence to find a hybridization match is dependent on its unique CoT value. The CoT value is a function of the number of copies (concentration) of the particular sequence, multiplied by the time of hybridization. It follows that for sequences which are abundant, hybridization events will occur very rapidly (low CoT value), while rare sequences will form duplexes at very high CoT values. CoT values which allow such rare sequences to form duplexes and therefore be effectively selected are difficult to achieve in a convenient time frame. Therefore, hybridization subtraction is simply not a useful technique with which to study relative levels of rare mRNA species. Sixth, this problem is further complicated by the fact that duplex formation is also dependent on the nucleotide base composition for a given sequence. Those sequences rich in G + C form stronger duplexes than those with high contents of A + T. Therefore, the former sequences will tend to be removed selectively by hybridization subtraction. Seventh, it is possible that hybridization between nonexact matches can occur. When this happens, the expression of a homologous gene may "mask" expression of a gene of interest, artificially skewing the results for that particular gene.

Matsubara and Okubo proposed using partial cDNA sequences to establish expression profiles of genes which could be used in functional analyses of the human genome. Matsubara and Okubo warned against using random priming, as

it creates multiple unique DNA fragments from individual mRNAs and may thus skew the analysis of the number of particular mRNAs per library. They sequenced randomly selected members from a 3'-directed cDNA library and established the frequency of appearance of the various ESTs. They proposed comparing lists of ESTs from various cell types to classify genes. Genes expressed in many different cell types were labeled housekeepers and those selectively expressed in certain cells were labeled cell-specific genes, even in the absence of the full sequence of the gene or the biological activity of the gene product.

The present invention avoids the drawbacks of the prior art by providing a method to quantify the relative abundance of multiple gene transcripts in a given biological specimen by the use of high-throughput sequence-specific analysis of individual RNAs and/or their corresponding cDNAs.

The present invention offers several advantages over current protein discovery methods which attempt to isolate individual proteins based upon biological effects. The method of the instant invention provides for detailed diagnostic comparisons of cell profiles revealing numerous changes in the expression of individual transcripts.

The instant invention provides several advantages over current subtraction methods including a more complex library analysis (10^6 to 10^7 clones as compared to 10^3 clones) which allows identification of low abundance messages as well as enabling the identification of messages which either increase or decrease in abundance. These large libraries are very routine to make in contrast to the libraries of previous methods. In addition, homologues can easily be distinguished with the method of the instant invention.

This method is very convenient because it organizes a large quantity of data into a comprehensible, digestible format. The most significant differences are highlighted by electronic subtraction. In depth analyses are made more convenient.

The present invention provides several advantages over previous methods of electronic analysis of cDNA. The method is particularly powerful when more than 100 and preferably more than 1,000 gene transcripts are analyzed. In such a case, new low-frequency transcripts are discovered and tissue typed.

High resolution analysis of gene expression can be used directly as a diagnostic profile or to identify disease-specific genes for the development of more classic diagnostic approaches.

This process is defined as gene transcript frequency analysis. The resulting quantitative analysis of the gene transcripts is defined as comparative gene transcript analysis.

3. SUMMARY OF THE INVENTION

The invention is a method of analyzing a specimen containing gene transcripts comprising the steps of (a) producing a library of biological sequences; (b) generating a set of transcript sequences, where each of the transcript sequences in said set is indicative of a different one of the biological sequences of the library; (c) processing the transcript sequences in a programmed computer (in which a database of reference transcript sequences indicative of reference sequences is stored), to generate an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of sequence annotation and a degree of match between one of the biological sequences of the library and at least one of the reference sequences; and (d) processing each said identified sequence value to generate final data values indicative of the number of times each identified sequence value is present in the library.

The invention also includes a method of comparing two specimens containing gene transcripts. The first specimen is processed as described above. The second specimen is used to produce a second library of biological sequences, which is used to generate a second set of transcript sequences, where each of the transcript sequences in the

second set is indicative of one of the biological sequences of the second library. Then the second set of transcript sequences is processed in a programmed computer to generate a second set of identified sequence values, namely the further identified sequence values, each of which is indicative of a sequence annotation and includes a degree of match between one of the biological sequences of the second library and at least one of the reference sequences. The further identified sequence values are processed to generate further final data values indicative of the number of times each further identified sequence value is present in the second library. The final data values from the first specimen and the further identified sequence values from the second specimen are processed to generate ratios of transcript sequences, which indicate the differences in the number of gene transcripts between the two specimens.

In a further embodiment, the method includes quantifying the relative abundance of mRNA in a biological specimen by (a) isolating a population of mRNA transcripts from a biological specimen; (b) identifying genes from which the mRNA was transcribed by a sequence-specific method; (c) determining the numbers of mRNA transcripts corresponding to each of the genes; and (d) using the mRNA transcript numbers to determine the relative abundance of mRNA transcripts within the population of mRNA transcripts.

Also disclosed is a method of producing a gene transcript image analysis by first obtaining a mixture of mRNA, from which cDNA copies are made. The cDNA is inserted into a suitable vector which is used to transfect suitable host strain cells which are plated out and permitted to grow into clones, each cone representing a unique mRNA. A representative population of clones transfected with cDNA is isolated. Each clone in the population is identified by a sequence-specific method which identifies the gene from which the unique mRNA was transcribed. The number of times each gene is identified to a clone is determined to evaluate gene transcript abundance. The genes and their abundances are listed in order of abundance to produce a gene transcript image.

In a further embodiment, the relative abundance of the gene transcripts in one cell type or tissue is compared with the relative abundance of gene transcript numbers in a second cell type or tissue in order to identify the differences and similarities.

In a further embodiment, the method includes a system for analyzing a library of biological sequences including a means for receiving a set of transcript sequences, where each of the transcript sequences is indicative of a different one of the biological sequences of the library; and a means for processing the transcript sequences in a computer system in which a database of reference transcript sequences indicative of reference sequences is stored, wherein the computer is programmed with software for generating an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence annotation and the degree of match between a different one of the biological sequences of the library and at least one of the reference sequences, and for processing each said identified sequence value to generate final data values indicative of the number of times each identified sequence value is present in the library.

In essence, the invention is a method and system for quantifying the relative abundance of gene transcripts in a biological specimen. The invention provides a method for comparing the gene transcript image from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes which are differentially expressed between the two specimens. Thus, this gene transcript image and its comparison can be used as a diagnostic. One embodiment of the method generates high-throughput sequence-specific analysis of multiple RNAs or their corresponding cDNAs: a gene transcript image. Another embodiment of the method produces the gene transcript imaging analysis by the use of high-throughput cDNA sequence analysis. In addition, two or more gene transcript images can be compared and used to detect or diagnose a particular biological state, disease,

or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells.

4. DESCRIPTION OF THE TABLES AND DRAWINGS

4.1. TABLES

Table 1 presents a detailed explanation of the letter codes utilized in Tables 2-5.

Table 2 lists the one hundred most common gene transcripts. It is a partial list of isolates from the HUVEC cDNA library prepared and sequenced as described below. The left-hand column refers to the sequence's order of abundance in this table. The next column labeled "number" is the clone number of the first HUVEC sequence identification reference matching the sequence in the "entry" column number. Isolates that have not been sequenced are not present in Table 2. The next column, labeled "N", indicates the total number of cDNAs which have the same degree of match with the sequence of the reference transcript in the "entry" column.

The column labeled "entry" gives the NIH GENBANK locus name, which corresponds to the library sequence numbers. The "s" column indicates in a few cases the species of the reference sequence. The code for column "s" is given in Table 1. The column labeled "descriptor" provides a plain English explanation of the identity of the sequence corresponding to the NIH GENBANK locus name in the "entry" column.

Table 3 is a comparison of the top fifteen most abundant gene transcripts in normal monocytes and activated macrophage cells.

Table 4 is a detailed summary of library subtraction analysis summary comparing the THP-1 and human macrophage cDNA sequences. In Table 4, the same code as in Table 2 is used. Additional columns are for "bgfreq" (abundance number in the subtractant library), "rfend" (abundance number in the target library) and "ratio" (the target abundance number divided by the subtractant abundance number). As is clear from perusal of the table, when the abundance number in the subtractant library is "0", the

target abundance number is divided by 0.05. This is a way of obtaining a result (not possible dividing by 0) and distinguishing the result from ratios of subtractant numbers of 1.

Table 5 is the computer program, written in source code, for generating gene transcript subtraction profiles.

Table 6 is a partial listing of database entries used in the electronic northern blot analysis as provided by the present invention.

4.2. BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a chart summarizing data collected and stored regarding the library construction portion of sequence preparation and analysis.

Figure 2 is a diagram representing the sequence of operations performed by "abundance sort" software in a class of preferred embodiments of the inventive method.

Figure 3 is a block diagram of a preferred embodiment of the system of the invention.

Figure 4 is a more detailed block diagram of the bioinformatics process from new sequence (that has already been sequenced but not identified) to printout of the transcript imaging analysis and the provision of database subscriptions.

5. DETAILED DESCRIPTION OF THE INVENTION

The present invention provides a method to compare the relative abundance of gene transcripts in different biological specimens by the use of high-throughput sequence-specific analysis of individual RNAs or their corresponding cDNAs (or alternatively, of data representing other biological sequences). This process is denoted herein as gene transcript imaging. The quantitative analysis of the relative abundance for a set of gene transcripts is denoted herein as "gene transcript image analysis" or "gene transcript frequency analysis". The present invention allows one to obtain a profile for gene transcription in any given population of cells or tissue from any type of organism. The invention can be applied to

obtain a profile of a specimen consisting of a single cell (or clones of a single cell), or of many cells, or of tissue more complex than a single cell and containing multiple cell types, such as liver.

The invention has significant advantages in the fields of diagnostics, toxicology and pharmacology, to name a few. A highly sophisticated diagnostic test can be performed on the ill patient in whom a diagnosis has not been made. A biological specimen consisting of the patient's fluids or tissues is obtained, and the gene transcripts are isolated and expanded to the extent necessary to determine their identity. Optionally, the gene transcripts can be converted to cDNA. A sampling of the gene transcripts are subjected to sequence-specific analysis and quantified. These gene transcript sequence abundances are compared against reference database sequence abundances including normal data sets for diseased and healthy patients. The patient has the disease(s) with which the patient's data set most closely correlates.

For example, gene transcript frequency analysis can be used to differentiate normal cells or tissues from diseased cells or tissues, just as it highlights differences between normal monocytes and activated macrophages in Table 3.

In toxicology, a fundamental question is which tests are most effective in predicting or detecting a toxic effect. Gene transcript imaging provides highly detailed information on the cell and tissue environment, some of which would not be obvious in conventional, less detailed screening methods. The gene transcript image is a more powerful method to predict drug toxicity and efficacy. Similar benefits accrue in the use of this tool in pharmacology. The gene transcript image can be used selectively to look at protein categories which are expected to be affected, for example, enzymes which detoxify toxins.

In an alternative embodiment, comparative gene transcript frequency analysis is used to differentiate between cancer cells which respond to anti-cancer agents and those which do not respond. Examples of anti-cancer

agents are tamoxifen, vincristine, vinblastine, podophyllotoxins, etoposide, tenisposide, cisplatin, biologic response modifiers such as interferon, Il-2, GM-CSF, enzymes, hormones and the like. This method also provides a means for sorting the gene transcripts by functional category. In the case of cancer cells, transcription factors or other essential regulatory molecules are very important categories to analyze across different libraries.

In yet another embodiment, comparative gene transcript frequency analysis is used to differentiate between control liver cells and liver cells isolated from patients treated with experimental drugs like FIAU to distinguish between pathology caused by the underlying disease and that caused by the drug.

In yet another embodiment, comparative gene transcript frequency analysis is used to differentiate between brain tissue from patients treated and untreated with lithium.

In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between cyclosporin and FK506-treated cells and normal cells.

In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between virally infected (including HIV-infected) human cells and uninfected human cells. Gene transcript frequency analysis is also used to rapidly survey gene transcripts in HIV-resistant, HIV-infected, and HIV-sensitive cells. Comparison of gene transcript abundance will indicate the success of treatment and/or new avenues to study.

In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between bronchial lavage fluids from healthy and unhealthy patients with a variety of ailments.

In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between cell, plant, microbial and animal mutants and wild-type species. In addition, the transcript abundance program is adapted to permit the scientist to evaluate the transcription of one gene in many different tissues. Such comparisons could

identify deletion mutants which do not produce a gene product and point mutants which produce a less abundant or otherwise different message. Such mutations can affect basic biochemical and pharmacological processes, such as mineral nutrition and metabolism, and can be isolated by means known to those skilled in the art. Thus, crops with improved yields, pest resistance and other factors can be developed.

In a further embodiment, comparative gene transcript frequency analysis is used for an interspecies comparative analysis which would allow for the selection of better pharmacologic animal models. In this embodiment, humans and other animals (such as a mouse), or their cultured cells are treated with a specific test agent. The relative sequence abundance of each cDNA population is determined. If the animal test system is a good model, homologous genes in the animal cDNA population should change expression similarly to those in human cells. If side effects are detected with the drug, a detailed transcript abundance analysis will be performed to survey gene transcript changes. Models will then be evaluated by comparing basic physiological changes.

In a further embodiment, comparative gene transcript frequency analysis is used in a clinical setting to give a highly detailed gene transcript profile of a patient's cells or tissue (for example, a blood sample). In particular, gene transcript frequency analysis is used to give a high resolution gene expression profile of a diseased state or condition.

In the preferred embodiment, the method utilizes high-throughput cDNA sequencing to identify specific transcripts of interest. The generated cDNA and deduced amino acid sequences are then extensively compared with GENBANK and other sequence data banks as described below. The method offers several advantages over current protein discovery by two-dimensional gel methods which try to identify individual proteins involved in a particular biological effect. Here, detailed comparisons of profiles of activated and inactive cells reveal numerous changes in

the expression of individual transcripts. After it is determined if the sequence is an "exact" match, similar or a non-match, the sequence is entered into a database. Next, the numbers of copies of cDNA corresponding to each gene are tabulated. Although this can be done slowly and arduously, if at all, by human hand from a printout of all entries, a computer program is a useful and rapid way to tabulate this information. The numbers of cDNA copies (optionally divided by the total number of sequences in the data set) provides a picture of the relative abundance of transcripts for each corresponding gene. The list of represented genes can then be sorted by abundance in the cDNA population. A multitude of additional types of comparisons or dimensions are possible and are exemplified below.

An alternate method of producing a gene transcript image includes the steps of obtaining a mixture of test mRNA and providing a representative array of unique probes whose sequences are complementary to at least some of the test mRNAs. Next, a fixed amount of the test mRNA is added to the arrayed probes. The test mRNA is incubated with the probes for a sufficient time to allow hybrids of the test mRNA and probes to form. The mRNA-probe hybrids are detected and the quantity determined. The hybrids are identified by their location in the probe array. The quantity of each hybrid is summed to give a population number. Each hybrid quantity is divided by the population number to provide a set of relative abundance data termed a gene transcript image analysis.

6. EXAMPLES

The examples below are provided to illustrate the subject invention. These examples are provided by way of illustration and are not included for the purpose of limiting the invention.

6.1. TISSUE SOURCES AND CELL LINES

For analysis with the computer program claimed herein, biological sequences can be obtained from virtually any

source. Most popular are tissues obtained from the human body. Tissues can be obtained from any organ of the body, any age donor, any abnormality or any immortalized cell line. Immortal cell lines may be preferred in some instances because of their purity of cell type; other tissue samples invariably include mixed cell types. A special technique is available to take a single cell (for example, a brain cell) and harness the cellular machinery to grow up sufficient cDNA for sequencing by the techniques and analysis described herein (cf. U.S. Patent Nos. 5,021,335 and 5,168,038, which are incorporated by reference). The examples given herein utilized the following immortalized cell lines: monocyte-like U-937 cells, activated macrophage-like THP-1 cells, induced vascular endothelial cells (HUVEC cells) and mast cell-like HMC-1 cells.

The U-937 cell line is a human histiocytic lymphoma cell line with monocyte characteristics, established from malignant cells obtained from the pleural effusion of a patient with diffuse histiocytic lymphoma (Sundstrom, C. and Nilsson, K. (1976) Int. J. Cancer 17:565). U-937 is one of only a few human cell lines with the morphology, cytochemistry, surface receptors and monocyte-like characteristics of histiocytic cells. These cells can be induced to terminal monocytic differentiation and will express new cell surface molecules when activated with supernatants from human mixed lymphocyte cultures. Upon this type of <u>in vitro</u> activation, the cells undergo morphological and functional changes, including augmentation of antibody-dependent cellular cytotoxicity (ADCC) against erythroid and tumor target cells (one of the principal functions of macrophages). Activation of U-937 cells with phorbol 12-myristate 13-acetate (PMA) <u>in vitro</u> stimulates the production of several compounds, including prostaglandins, leukotrienes and platelet-activating factor (PAF), which are potent inflammatory mediators. Thus, U-937 is a cell line that is well suited for the identification and isolation of gene transcripts associated with normal monocytes.

The HUVEC cell line is a normal, homogeneous, well characterized, early passage endothelial cell culture from human umbilical vein (Cell Systems Corp., 12815 NE 124th Street, Kirkland, WA 98034). Only gene transcripts from induced, or treated, HUVEC cells were sequenced. One batch of 1 X 10^8 cells was treated for 5 hours with 1 U/ml rIL-1b and 100 ng/ml E.coli lipopolysaccharide (LPS) endotoxin prior to harvesting. A separate batch of 2 X 10^8 cells was treated at confluence with 4 U/ml TNF and 2 U/ml interferon-gamma (IFN-gamma) prior to harvesting.

THP-1 is a human leukemic cell line with distinct monocytic characteristics. This cell line was derived from the blood of a 1-year-old boy with acute monocytic leukemia (Tsuchiya, S. et al. (1980) Int. J. Cancer: 171-76). The following cytological and cytochemical criteria were used to determine the monocytic nature of the cell line: 1) the presence of alpha-naphthyl butyrate esterase activity which could be inhibited by sodium fluoride; 2) the production of lysozyme; 3) the phagocytosis of latex particles and sensitized SRBC (sheep red blood cells); and 4) the ability of mitomycin C-treated THP-1 cells to activate T-lymphocytes following ConA (concanavalin A) treatment. Morphologically, the cytoplasm contained small azurophilic granules and the nucleus was indented and irregularly shaped with deep folds. The cell line had Fc and C3b receptors, probably functioning in phagocytosis. THP-1 cells treated with the tumor promoter 12-o-tetradecanoyl-phorbol-13 acetate (TPA) stop proliferating and differentiate into macrophage-like cells which mimic native monocyte-derived macrophages in several respects. Morphologically, as the cells change shape, the nucleus becomes more irregular and additional phagocytic vacuoles appear in the cytoplasm. The differentiated THP-1 cells also exhibit an increased adherence to tissue culture plastic.

HMC-1 cells (a human mast cell line) were established from the peripheral blood of a Mayo Clinic patient with mast cell leukemia (Leukemia Res. (1988) 12:345-55). The cultured cells looked similar to immature cloned murine

mast cells, contained histamine, and stained positively for chloroacetate esterase, amino caproate esterase, eosinophil major basic protein (MBP) and tryptase. The HMC-1 cells have, however, lost the ability to synthesize normal IgE receptors. HMC-1 cells also possess a 10;16 translocation, present in cells initially collected by leukophoresis from the patient and not an artifact of culturing. Thus, HMC-1 cells are a good model for mast cells.

6.2. CONSTRUCTION OF cDNA LIBRARIES

For inter-library comparisons, the libraries must be prepared in similar manners. Certain parameters appear to be particularly important to control. One such parameter is the method of isolating mRNA. It is important to use the same conditions to remove DNA and heterogeneous nuclear RNA from comparison libraries. Size fractionation of cDNA must be carefully controlled. The same vector preferably should be used for preparing libraries to be compared. At the very least, the same type of vector (e.g., unidirectional vector) should be used to assure a valid comparison. A unidirectional vector may be preferred in order to more easily analyze the output.

It is preferred to prime only with oligo dT unidirectional primer in order to obtain one only clone per mRNA transcript when obtaining cDNAs. However, it is recognized that employing a mixture of oligo dT and random primers can also be advantageous because such a mixture results in more sequence diversity when gene discovery also is a goal. Similar effects can be obtained with DR2 (Clontech) and HXLOX (US Biochemical) and also vectors from Invitrogen and Novagen. These vectors have two requirements. First, there must be primer sites for commercially available primers such as T3 or M13 reverse primers. Second, the vector must accept inserts up to 10 kB.

It also is important that the clones be randomly sampled, and that a significant population of clones is used. Data have been generated with 5,000 clones; however, if very rare genes are to be obtained and/or their relative

abundance determined, as many as 100,000 clones from a single library may need to be sampled. Size fractionation of cDNA also must be carefully controlled. Alternately, plaques can be selected, rather than clones.

Besides the Uni-ZAP™ vector system by Stratagene disclosed below, it is now believed that other similarly unidirectional vectors also can be used. For example, it is believed that such vectors include but are not limited to DR2 (Clontech), and HXLOX (U.S. Biochemical).

Preferably, the details of library construction (as shown in Figure 1) are collected and stored in a database for later retrieval relative to the sequences being compared. Fig. 1 shows important information regarding the library collaborator or cell or cDNA supplier, pretreatment, biological source, culture, mRNA preparation and cDNA construction. Similarly detailed information about the other steps is beneficial in analyzing sequences and libraries in depth.

RNA must be harvested from cells and tissue samples and cDNA libraries are subsequently constructed. cDNA libraries can be constructed according to techniques known in the art. (See, for example, Maniatis, T. et al. (1982) Molecular Cloning, Cold Spring Harbor Laboratory, New York). cDNA libraries may also be purchased. The U-937 cDNA library (catalog No. 937207) was obtained from Stratagene, Inc., 11099 M. Torrey Pines Rd., La Jolla, CA 92037.

The THP-1 cDNA library was custom constructed by Stratagene from THP-1 cells cultured 48 hours with 100 nm TPA and 4 hours with 1 μg/ml LPS. The human mast cell HMC-1 cDNA library was also custom constructed by Stratagene from cultured HMC-1 cells. The HUVEC cDNA library was custom constructed by Stratagene from two batches of induced HUVEC cells which were separately processed.

Essentially, all the libraries were prepared in the same manner. First, poly(A+)RNA (mRNA) was purified. For the U-937 and HMC-1 RNA, cDNA synthesis was only primed with oligo dT. For the THP-1 and HUVEC RNA, cDNA synthesis was primed separately with both oligo dT and random

hexamers, and the two cDNA libraries were treated separately. Synthetic adaptor oligonucleotides were ligated onto cDNA ends enabling its insertion into the Uni-Zap™ vector system (Stratagene), allowing high efficiency unidirectional (sense orientation) lambda library construction and the convenience of a plasmid system with blue-white color selection to detect clones with cDNA insertions. Finally, the two libraries were combined into a single library by mixing equal numbers of bacteriophage.

The libraries can be screened with either DNA probes or antibody probes and the pBluescript® phagemid (Stratagene) can be rapidly excised in vivo. The phagemid allows the use of a plasmid system for easy insert characterization, sequencing, site-directed mutagenesis, the creation of unidirectional deletions and expression of fusion proteins. The custom-constructed library phage particles were infected into E. coli host strain XL1-Blue® (Stratagene), which has a high transformation efficiency, increasing the probability of obtaining rare, under-represented clones in the cDNA library.

6.3. ISOLATION OF cDNA CLONES

The phagemid forms of individual cDNA clones were obtained by the in vivo excision process, in which the host bacterial strain was coinfected with both the lambda library phage and an f1 helper phage. Proteins derived from both the library-containing phage and the helper phage nicked the lambda DNA, initiated new DNA synthesis from defined sequences on the lambda target DNA and created a smaller, single stranded circular phagemid DNA molecule that included all DNA sequences of the pBluescript® plasmid and the cDNA insert. The phagemid DNA was secreted from the cells and purified, then used to re-infect fresh host cells, where the double stranded phagemid DNA was produced. Because the phagemid carries the gene for beta-lactamase, the newly-transformed bacteria are selected on medium containing ampicillin.

Phagemid DNA was purified using the Magic Minipreps™ DNA Purification System (Promega catalogue #A7100. Promega

Corp., 2800 Woods Hollow Rd., Madison, WI 53711). This small-scale process provides a simple and reliable method for lysing the bacterial cells and rapidly isolating purified phagemid DNA using a proprietary DNA-binding resin. The DNA was eluted from the purification resin already prepared for DNA sequencing and other analytical manipulations.

Phagemid DNA was also purified using the QIAwell-8 Plasmid Purification System from QIAGEN® DNA Purification System (QIAGEN Inc., 9259 Eton Ave., Chattsworth, CA 91311). This product line provides a convenient, rapid and reliable high-throughput method for lysing the bacterial cells and isolating highly purified phagemid DNA using QIAGEN anion-exchange resin particles with EMPORE™ membrane technology from 3M in a multiwell format. The DNA was eluted from the purification resin already prepared for DNA sequencing and other analytical manipulations.

An alternate method of purifying phagemid has recently become available. It utilizes the Miniprep Kit (Catalog No. 77468, available from Advanced Genetic Technologies Corp., 19212 Orbit Drive, Gaithersburg, Maryland). This kit is in the 96-well format and provides enough reagents for 960 purifications. Each kit is provided with a recommended protocol, which has been employed except for the following changes. First, the 96 wells are each filled with only 1 ml of sterile terrific broth with carbenicillin at 25 mg/L and glycerol at 0.4%. After the wells are inoculated, the bacteria are cultured for 24 hours and lysed with 60 µl of lysis buffer. A centrifugation step (2900 rpm for 5 minutes) is performed before the contents of the block are added to the primary filter plate. The optional step of adding isopropanol to TRIS buffer is not routinely performed. After the last step in the protocol, samples are transferred to a Beckman 96-well block for storage.

Another new DNA purification system is the WIZARD™ product line which is available from Promega (catalog No. A7071) and may be adaptable to the 96-well format.

6.4. SEQUENCING OF cDNA CLONES

The cDNA inserts from random isolates of the U-937 and THP-1 libraries were sequenced in part. Methods for DNA sequencing are well known in the art. Conventional enzymatic methods employ DNA polymerase Klenow fragment, Sequenase™ or Taq polymerase to extend DNA chains from an oligonucleotide primer annealed to the DNA template of interest. Methods have been developed for the use of both single- and double-stranded templates. The chain termination reaction products are usually electrophoresed on urea-acrylamide gels and are detected either by autoradiography (for radionuclide-labeled precursors) or by fluorescence (for fluorescent-labeled precursors). Recent improvements in mechanized reaction preparation, sequencing and analysis using the fluorescent detection method have permitted expansion in the number of sequences that can be determined per day (such as the Applied Biosystems 373 and 377 DNA sequencer, Catalyst 800). Currently with the system as described, read lengths range from 250 to 400 bases and are clone dependent. Read length also varies with the length of time the gel is run. In general, the shorter runs tend to truncate the sequence. A minimum of only about 25 to 50 bases is necessary to establish the identification and degree of homology of the sequence. Gene transcript imaging can be used with any sequence-specific method, including, but not limited to hybridization, mass spectroscopy, capillary electrophoresis and 505 gel electrophoresis.

6.5. HOMOLOGY SEARCHING OF cDNA CLONE AND DEDUCED PROTEIN (and Subsequent Steps)

Using the nucleotide sequences derived from the cDNA clones as query sequences (sequences of a Sequence Listing), databases containing previously identified sequences are searched for areas of homology (similarity). Examples of such databases include Genbank and EMBL. We next describe examples of two homology search algorithms that can be used, and then describe the subsequent computer-implemented steps to be performed in accordance with preferred embodiments of the invention.

In the following description of the computer-implemented steps of the invention, the word "library" denotes a set (or population) of biological specimen nucleic acid sequences. A "library" can consist of cDNA sequences, RNA sequences, or the like, which characterize a biological specimen. The biological specimen can consist of cells of a single human cell type (or can be any of the other above-mentioned types of specimens). We contemplate that the sequences in a library have been determined so as to accurately represent or characterize a biological specimen (for example, they can consist of representative cDNA sequences from clones of RNA taken from a single human cell).

In the following description of the computer-implemented steps of the invention, the expression "database" denotes a set of stored data which represent a collection of sequences, which in turn represent a collection of biological reference materials. For example, a database can consist of data representing many stored cDNA sequences which are in turn representative of human cells infected with various viruses, cells of humans of various ages, cells from different mammalian species, and so on.

In preferred embodiments, the invention employs a computer programmed with software (to be described) for performing the following steps:

(a) processing data indicative of a library of cDNA sequences (generated as a result of high-throughput cDNA sequencing or other method) to determine whether each sequence in the library matches a DNA sequence of a reference database of DNA sequences (and if so, identifying the reference database entry which matches the sequence and indicating the degree of match between the reference sequence and the library sequence) and assigning an identified sequence value based on the sequence annotation and degree of match to each of the sequences in the library;

(b) for some or all entries of the database, tabulating the number of matching identified sequence

values in the library (Although this can be done by human hand from a printout of all entries, we prefer to perform this step using computer software to be described below.), thereby generating a set of final data values or "abundance numbers"; and

(c) if the libraries are different sizes, dividing each abundance number by the total number of sequences in the library, to obtain a relative abundance number for each identified sequence value (i.e., a relative abundance of each gene transcript).

The list of identified sequence values (or genes corresponding thereto) can then be sorted by abundance in the cDNA population. A multitude of additional types of comparisons or dimensions are possible.

For example (to be described below in greater detail), steps (a) and (b) can be repeated for two different libraries (sometimes referred to as a "target" library and a "subtractant" library). Then, for each identified sequence value (or gene transcript), a "ratio" value is obtained by dividing the abundance number (for that identified sequence value) for the target library, by the abundance number (for that identified sequence value) for the subtractant library.

In fact, subtraction may be carried out on multiple libraries. It is possible to add the transcripts from several libraries (for example, three) and then to divide them by another set of transcripts from multiple libraries (again, for example, three). Notation for this operation may be abbreviated as (A+B+C) / (D+E+F), where the capital letters each indicate an entire library. Optionally the abundance numbers of transcripts in the summed libraries may be divided by the total sample size before subtraction.

Unlike standard hybridization technology which permits a single subtraction of two libraries, once one has processed a set or library transcript sequences and stored them in the computer, any number of subtractions can be performed on the library. For example, by this method, ratio values can be obtained by dividing relative abundance

values in a first library by corresponding values in a second library and vice versa.

In variations on step (a), the library consists of nucleotide sequences derived from cDNA clones. Examples of databases which can be searched for areas of homology (similarity) in step (a) include the commercially available databases known as Genbank (NIH) EMBL (European Molecular Biology Labs, Germany), and GENESEQ (Intelligenetics, Mountain View, California).

One homology search algorithm which can be used to implement step (a) is the algorithm described in the paper by D.J. Lipman and W.R. Pearson, entitled "Rapid and Sensitive Protein Similarity Searches," Science, 227:1435 (1985). In this algorithm, the homologous regions are searched in a two-step manner. In the first step, the highest homologous regions are determined by calculating a matching score using a homology score table. The parameter "Ktup" is used in this step to establish the minimum window size to be shifted for comparing two sequences. Ktup also sets the number of bases that must match to extract the highest homologous region among the sequences. In this step, no insertions or deletions are applied and the homology is displayed as an initial (INIT) value.

In the second step, the homologous regions are aligned to obtain the highest matching score by inserting a gap in order to add a probable deleted portion. The matching score obtained in the first step is recalculated using the homology score Table and the insertion score Table to an optimized (OPT) value in the final output.

DNA homologies between two sequences can be examined graphically using the Harr method of constructing dot matrix homology plots (Needleman, S.B. and Wunsch, C.O., J. Mom. Biol 48:443 (1970)). This method produces a two-dimensional plot which can be useful in determining regions of homology versus regions of repetition.

However, in a class of preferred embodiments, step (a) is implemented by processing the library data in the commercially available computer program known as the INHERIT 670 Sequence Analysis System, available from

Applied Biosystems Inc. (Foster City, California), including the software known as the Factura software (also available from Applied Biosystems Inc.). The Factura program preprocesses each library sequence to "edit out" portions thereof which are not likely to be of interest, such as the vector used to prepare the library. Additional sequences which can be edited out or masked (ignored by the search tools) include but are not limited to the polyA tail and repetitive GAG and CCC sequences. A low-end search program can be written to mask out such "low-information" sequences, or programs such as BLAST can ignore the low-information sequences.

In the algorithm implemented by the INHERIT 670 Sequence Analysis System, the Pattern Specification Language (developed by TRW Inc.) is used to determine regions of homology. "There are three parameters that determine how INHERIT analysis runs sequence comparisons: window size, window offset and error tolerance. Window size specifies the length of the segments into which the query sequence is subdivided. Window offset specifies where to start the next segment [to be compared], counting from the beginning of the previous segment. Error tolerance specifies the total number of insertions, deletions and/or substitutions that are tolerated over the specified word length. Error tolerance may be set to any integer between 0 and 6. The default settings are window tolerance=20, window offset=10 and error tolerance=3." INHERIT Analysis Users Manual, pp.2-15. Version 1.0, Applied Biosystems, Inc., October 1991.

Using a combination of these three parameters, a database (such as a DNA database) can be searched for sequences containing regions of homology and the appropriate sequences are scored with an initial value. Subsequently, these homologous regions are examined using dot matrix homology plots to determine regions of homology versus regions of repetition. Smith-Waterman alignments can be used to display the results of the homology search. The INHERIT software can be executed by a Sun computer system programmed with the UNIX operating system.

Search alternatives to INHERIT include the BLAST program, GCG (available from the Genetics Computer Group, WI) and the Dasher program (Temple Smith, Boston University, Boston, MA). Nucleotide sequences can be searched against Genbank, EMBL or custom databases such as GENESEQ (available from Intelligenetics, Mountain View, CA) or other databases for genes. In addition, we have searched some sequences against our own in-house database.

In preferred embodiments, the transcript sequences are analyzed by the INHERIT software for best conformance with a reference gene transcript to assign a sequence identifier and assigned the degree of homology, which together are the identified sequence value and are input into, and further processed by, a Macintosh personal computer (available from Apple) programmed with an "abundance sort and subtraction analysis" computer program (to be described below).

Prior to the abundance sort and subtraction analysis program (also denoted as the "abundance sort" program), identified sequences from the cDNA clones are assigned value (according to the parameters given above) by degree of match according to the following categories: "exact" matches (regions with a high degree of identity), homologous human matches (regions of high similarity, but not "exact" matches), homologous non-human matches (regions of high similarity present in species other than human), or non matches (no significant regions of homology to previously identified nucleotide sequences stored in the form of the database). Alternately, the degree of match can be a numeric value as described below.

With reference again to the step of identifying matches between reference sequences and database entries, protein and peptide sequences can be deduced from the nucleic acid sequences. Using the deduced polypeptide sequence, the match identification can be performed in a manner analogous to that done with cDNA sequences. A protein sequence is used as a query sequence and compared to the previously identified sequences contained in a database such as the Swiss/Prot, PIR and the NBRF Protein database to find homologous proteins. These proteins are

initially scored for homology using a homology score Table (Orcutt, B.C. and Dayoff, M.O. Scoring Matrices, PIR Report MAT - 0285 (February 1985)) resulting in an INIT score. The homologous regions are aligned to obtain the highest matching scores by inserting a gap which adds a probable deleted portion. The matching score is recalculated using the homology score Table and the insertion score Table resulting in an optimized (OPT) score. Even in the absence of knowledge of the proper reading frame of an isolated sequence, the above-described protein homology search may be performed by searching all 3 reading frames.

Peptide and protein sequence homologies can also be ascertained using the INHERIT 670 Sequence Analysis System in an analogous way to that used in DNA sequence homologies. Pattern Specification Language and parameter windows are used to search protein databases for sequences containing regions of homology which are scored with an initial value. Subsequent display in a dot-matrix homology plot shows regions of homology versus regions of repetition. Additional search tools that are available to use on pattern search databases include PLsearch Blocks (available from Henikoff & Henikoff, University of Washington, Seattle), Dasher and GCG. Pattern search databases include, but are not limited to, Protein Blocks (available from Henikoff & Henikoff, University of Washington, Seattle), Brookhaven Protein (available from the Brookhaven National Laboratory, Brookhaven, MA), PROSITE (available from Amos Bairoch, University of Geneva, Switzerland), ProDom (available from Temple Smith, Boston University), and PROTEIN MOTIF FINGERPRINT (available from University of Leeds, United Kingdom).

The ABI Assembler application software, part of the INHERIT DNA analysis system (available from Applied Biosystems, Inc., Foster City, CA), can be employed to create and manage sequence assembly projects by assembling data from selected sequence fragments into a larger sequence. The Assembler software combines two advanced computer technologies which maximize the ability to

assemble sequenced DNA fragments into Assemblages, a special grouping of data where the relationships between sequences are shown by graphic overlap, alignment and statistical views. The process is based on the Meyers-Kececioglu model of fragment assembly (INHERIT™ Assembler User's Manual, Applied Biosystems, Inc., Foster City, CA), and uses graph theory as the foundation of a very rigorous multiple sequence alignment engine for assembling DNA sequence fragments. Other assembly programs that can be used include MEGALIGN (available from DNASTAR Inc., Madison, WI), Dasher and STADEN (available from Roger Staden, Cambridge, England).

Next, with reference to Fig. 2, we describe in more detail the "abundance sort" program which implements above-mentioned "step (b)" to tabulate the number of sequences of the library which match each database entry (the "abundance number" for each database entry).

Fig. 2 is a flow chart of a preferred embodiment of the abundance sort program. A source code listing of this embodiment of the abundance sort program is set forth in Table 5. In the Table 5 implementation, the abundance sort program is written using the FoxBASE programming language commercially available from Microsoft Corporation. Although FoxBASE was the program chosen for the first iteration of this technology, it should not be considered limiting. Many other programming languages, Sybase being a particularly desirable alternative, can also be used, as will be obvious to one with ordinary skill in the art. The subroutine names specified in Fig. 2 correspond to subroutines listed in Table 5.

With reference again to Fig. 2, the "Identified Sequences" are transcript sequences representing each sequence of the library and a corresponding identification of the database entry (if any) which it matches. In other words, the "Identified Sequences" are transcript sequences representing the output of above-discussed "step (a)."

Fig. 3 is a block diagram of a system for implementing the invention. The Fig. 3 system includes library generation unit 2 which generates a library and asserts an

output stream of transcript sequences indicative of the biological sequences comprising the library. Programmed processor 4 receives the data stream output from unit 2 and processes this data in accordance with above-discussed "step (a)" to generate the Identified Sequences. Processor 4 can be a processor programmed with the commercially available computer program known as the INHERIT 670 Sequence Analysis System and the commercially available computer program known as the Factura program (both available from Applied Biosystems Inc.) and with the UNIX operating system.

Still with reference to Fig. 3, the Identified Sequences are loaded into processor 6 which is programmed with the abundance sort program. Processor 6 generates the Final Transcript sequences indicated in both Figs. 2 and 3. Fig. 4 shows a more detailed block diagram of a planned relational computer system, including various searching techniques which can be implemented, along with an assortment of databases to query against.

With reference to Fig. 2, the abundance sort program first performs an operation known as "Tempnum" on the Identified Sequences, to discard all of the Identified Sequences except those which match database entries of selected types. For example, the Tempnum process can select Identified Sequences which represent matches of the following types with database entries (see above for definition): "exact" matches, human "homologous" matches, "other species" matches representing genes present in species other than human), "no" matches (no significant regions of homology with database entries representing previously identified nucleotide sequences), "I" matches (Incyte for not previously known DNA sequences), or "X" matches (matches ESTs in reference database). This eliminates the U, S, M, V, A, R and D sequence (see Table 1 for definitions).

The identified sequence values selected during the "Tempnum" process then undergo a further selection (weeding out) operation known as "Tempred." This operation can, for

example, discard all identified sequence values representing matches with selected database entries.

The identified sequence values selected during the "Tempred" process are then classified according to library, during the "Tempdesig" operation. It is contemplated that the "Identified Sequences" can represent sequences from a single library, or from two or more libraries.

Consider first the case that the identified sequence values represent sequences from a single library. In this case, all the identified sequence values determined during "Tempred" undergo sorting in the "Templib" operation, further sorting in the "Libsort" operation, and finally additional sorting in the "Temptarsort" operation. For example, these three sorting operations can sort the identified sequences in order of decreasing "abundance number" (to generate a list of decreasing abundance numbers, each abundance number corresponding to a unique identified sequence entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected type) with redundancies eliminated from each sorted list. In this case, the operation identified as "Cruncher" can be bypassed, so that the "Final Data" values are the organized transcript sequences produced during the "Temptarsort" operation.

We next consider the case that the transcript sequences produced during the "Tempred" operation represent sequences from two libraries (which we will denote the "target" library and the "subtractant" library). For example, the target library may consist of cDNA sequences from clones of a diseased cell, while the subtractant library may consist of cDNA sequences from clones of the diseased cell after treatment by exposure to a drug. For another example, the target library may consist of cDNA sequences from clones of a cell type from a young human, while the subtractant library may consist of cDNA sequences from clones of the same cell type from the same human at different ages.

In this case, the "Tempdesig" operation routes all transcript sequences representing the target library for processing in accordance with "Templib" (and then "Libsort" and "Temptarsort"), and routes all transcript sequences representing the subtractant library for processing in accordance with "Tempsub" (and then "Subsort" and "Tempsubsort"). For example, the consecutive "Templib," "Libsort," and "Temptarsort" sorting operations sort identified sequences from the target library in order of decreasing abundance number (to generate a list of decreasing abundance numbers, each abundance number corresponding to a database entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected type) with redundancies eliminated from each sorted list. The consecutive "Tempsub," "Subsort," and "Tempsubsort" sorting operations sort identified sequences from the subtractant library in order of decreasing abundance number (to generate a list of decreasing abundance numbers, each abundance number corresponding to a database entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected type) with redundancies eliminated from each sorted list.

The transcript sequences output from the "Temptarsort" operation typically represent sorted lists from which a histogram could be generated in which position along one (e.g., horizontal) axis indicates abundance number (of target library sequences), and position along another (e.g., vertical) axis indicates identified sequence value (e.g., human or non-human gene type). Similarly, the transcript sequences output from the "Tempsubsort" operation typically represent sorted lists from which a histogram could be generated in which position along one (e.g., horizontal) axis indicates abundance number (of subtractant library sequences), and position along another (e.g., vertical) axis indicates identified sequence value (e.g., human or non-human gene type).

The transcript sequences (sorted lists) output from the Tempsubsort and Temptarsort sorting operations are combined during the operation identified as "Cruncher." The "Cruncher" process identifies pairs of corresponding target and subtractant abundance numbers (both representing the same identified sequence value), and divides one by the other to generate a "ratio" value for each pair of corresponding abundance numbers, and then sorts the ratio values in order of decreasing ratio value. The data output from the "Cruncher" operation (the Final Transcript sequence in Fig. 2) is typically a sorted list from which a histogram could be generated in which position along one axis indicates the size of a ratio of abundance numbers (for corresponding identified sequence values from target and subtractant libraries) and position along another axis indicates identified sequence value (e.g., gene type).

Preferably, prior to obtaining a ratio between the two library abundance values, the Cruncher operation also divides each ratio value by the total number of sequences in one or both of the target and subtractant libraries. The resulting lists of "relative" ratio values generated by the Cruncher operation are useful for many medical, scientific, and industrial applications. Also preferably, the output of the Cruncher operation is a set of lists, each list representing a sequence of decreasing ratio values for a different selected subset (e.g. protein family) of database entries.

In one example, the abundance sort program of the invention tabulates for a library the numbers of mRNA transcripts corresponding to each gene identified in a database. These numbers are divided by the total number of clones sampled. The results of the division reflect the relative abundance of the mRNA transcripts in the cell type or tissue from which they were obtained. Obtaining this final data set is referred to herein as "gene transcript image analysis." The resulting subtracted data show exactly what proteins and genes are upregulated and downregulated in highly detailed complexity.

6.6. HUVEC cDNA LIBRARY

Table 2 is an abundance table listing the various gene transcripts in an induced HUVEC library. The transcripts are listed in order of decreasing abundance. This computerized sorting simplifies analysis of the tissue and speeds identification of significant new proteins which are specific to this cell type. This type of endothelial cell lines tissues of the cardiovascular system, and the more that is known about its composition, particularly in response to activation, the more choices of protein targets become available to affect in treating disorders of this tissue, such as the highly prevalent atherosclerosis.

6.7. MONOCYTE-CELL AND MAST-CELL cDNA LIBRARIES

Tables 3 and 4 show truncated comparisons of two libraries. In Tables 3 and 4 the "normal monocytes" are the HMC-1 cells, and the "activated macrophages" are the THP-1 cells pretreated with PMA and activated with LPS. Table 3 lists in descending order of abundance the most abundant gene transcripts for both cell types. With only 15 gene transcripts from each cell type, this table permits quick, qualitative comparison of the most common transcripts. This abundance sort, with its convenient side-by-side display, provides an immediately useful research tool. In this example, this research tool discloses that 1) only one of the top 15 activated macrophage transcripts is found in the top 15 normal monocyte gene transcripts (poly A binding protein); and 2) a new gene transcript (previously unreported in other databases) is relatively highly represented in activated macrophages but is not similarly prominent in normal macrophages. Such a research tool provides researchers with a short-cut to new proteins, such as receptors, cell-surface and intracellular signalling molecules, which can serve as drug targets in commercial drug screening programs. Such a tool could save considerable time over that consumed by a hit and miss discovery program aimed at identifying important proteins in and around cells, because those proteins carrying out everyday cellular functions and

represented as steady state mRNA are quickly eliminated from further characterization.

This illustrates how the gene transcript profiles change with altered cellular function. Those skilled in the art know that the biochemical composition of cells also changes with other functional changes such as cancer, including cancer's various stages, and exposure to toxicity. A gene transcript subtraction profile such as in Table 3 is useful as a first screening tool for such gene expression and protein studies.

6.8. SUBTRACTION ANALYSIS OF NORMAL MONOCYTE-CELL AND ACTIVATED MONOCYTE CELL cDNA LIBRARIES

Once the cDNA data are in the computer, the computer program as disclosed in Table 5 was used to obtain ratios of all the gene transcripts in the two libraries discussed in Example 6.7, and the gene transcripts were sorted by the descending values of their ratios. If a gene transcript is not represented in one library, that gene transcript's abundance is unknown but appears to be less than 1. As an approximation -- and to obtain a ratio, which would not be possible if the unrepresented gene were given an abundance of zero -- genes which are represented in only one of the two libraries are assigned an abundance of 1/2. Using 1/2 for unrepresented clones increases the relative importance of "turned-on" and "turned-off" genes, whose products would be drug candidates. The resulting print-out is called a subtraction table and is an extremely valuable screening method, as is shown by the following data.

Table 4 is a subtraction table, in which the normal monocyte library was electronically "subtracted" from the activated macrophage library. This table highlights most effectively the changes in abundance of the gene transcripts by activation of macrophages. Even among the first 20 gene transcripts listed, there are several unknown gene transcripts. Thus, electronic subtraction is a useful tool with which to assist researchers in identifying much more quickly the basic biochemical changes between two cell types. Such a tool can save universities and pharmaceutical companies which spend billions of dollars on

research valuable time and laboratory resources at the early discovery stage and can speed up the drug development cycle, which in turn permits researchers to set up drug screening programs much earlier. Thus, this research tool provides a way to get new drugs to the public faster and more economically.

Also, such a subtraction table can be obtained for patient diagnosis. An individual patient sample (such as monocytes obtained from a biopsy or blood sample) can be compared with data provided herein to diagnose conditions associated with macrophage activation.

Table 4 uncovered many new gene transcripts (labeled Incyte clones). Note that many genes are turned on in the activated macrophage (i.e., the monocyte had a 0 in the bgfreq column). This screening method is superior to other screening techniques, such as the western blot, which are incapable of uncovering such a multitude of discrete new gene transcripts.

The subtraction-screening technique has also uncovered a high number of cancer gene transcripts (oncogenes rho, ETS2, rab-2 ras, YPT1-related, and acute myeloid leukemia mRNA) in the activated macrophage. These transcripts may be attributed to the use of immortalized cell lines and are inherently interesting for that reason. This screening technique offers a detailed picture of upregulated transcripts including oncogenes, which helps explain why anti-cancer drugs interfere with the patient's immunity mediated by activated macrophages. Armed with knowledge gained from this screening method, those skilled in the art can set up more targeted, more effective drug screening programs to identify drugs which are differentially effective against 1) both relevant cancers and activated macrophage conditions with the same gene transcript profile; 2) cancer alone; and 3) activated macrophage conditions.

Smooth muscle senescent protein (22 kd) was upregulated in the activated macrophage, which indicates that it is a candidate to block in controlling inflammation.

6.9. SUBTRACTION ANALYSIS OF NORMAL LIVER CELLS AND HEPATITIS INFECTED LIVER CELL cDNA LIBRARIES

In this example, rats are exposed to hepatitis virus and maintained in the colony until they show definite signs of hepatitis. Of the rats diagnosed with hepatitis, one half of the rats are treated with a new anti-hepatitis agent (AHA). Liver samples are obtained from all rats before exposure to the hepatitis virus and at the end of AHA treatment or no treatment. In addition, liver samples can be obtained from rats with hepatitis just prior to AHA treatment.

The liver tissue is treated as described in Examples 6.2 and 6.3 to obtain mRNA and subsequently to sequence cDNA. The cDNA from each sample are processed and analyzed for abundance according to the computer program in Table 5. The resulting gene transcript images of the cDNA provide detailed pictures of the baseline (control) for each animal and of the infected and/or treated state of the animals. cDNA data for a group of samples can be combined into a group summary gene transcript profile for all control samples, all samples from infected rats and all samples from AHA-treated rats.

Subtractions are performed between appropriate individual libraries and the grouped libraries. For individual animals, control and post-study samples can be subtracted. Also, if samples are obtained before and after AHA treatment, that data from individual animals and treatment groups can be subtracted. In addition, the data for all control samples can be pooled and averaged. The control average can be subtracted from averages of both post-study AHA and post-study non-AHA cDNA samples. If pre- and post-treatment samples are available, pre- and post-treatment samples can be compared individually (or electronically averaged) and subtracted.

These subtraction tables are used in two general ways. First, the differences are analyzed for gene transcripts which are associated with continuing hepatic deterioration or healing. The subtraction tables are tools to isolate the effects of the drug treatment from the underlying basic pathology of hepatitis. Because hepatitis affects many

parameters, additional liver toxicity has been difficult to detect with only blood tests for the usual enzymes. The gene transcript profile and subtraction provides a much more complex biochemical picture which researchers have needed to analyze such difficult problems.

Second, the subtraction tables provide a tool for identifying clinical markers, individual proteins or other biochemical determinants which are used to predict and/or evaluate a clinical endpoint, such as disease, improvement due to the drug, and even additional pathology due to the drug. The subtraction tables specifically highlight genes which are turned on or off. Thus, the subtraction tables provide a first screen for a set of gene transcript candidates for use as clinical markers. Subsequently, electronic subtractions of additional cell and tissue libraries reveal which of the potential markers are in fact found in different cell and tissue libraries. Candidate gene transcripts found in additional libraries are removed from the set of potential clinical markers. Then, tests of blood or other relevant samples which are known to lack and have the relevant condition are compared to validate the selection of the clinical marker. In this method, the particular physiologic function of the protein transcript need not be determined to qualify the gene transcript as a clinical marker.

6.10. ELECTRONIC NORTHERN BLOT

One limitation of electronic subtraction is that it is difficult to compare more than a pair of images at once. Once particular individual gene products are identified as relevant to further study (via electronic subtraction or other methods), it is useful to study the expression of single genes in a multitude of different tissues. In the lab, the technique of "Northern" blot hybridization is used for this purpose. In this technique, a single cDNA, or a probe corresponding thereto, is labeled and then hybridized against a blot containing RNA samples prepared from a multitude of tissues or cell types. Upon autoradiography,

the pattern of expression of that particular gene, one at a time, can be quantitated in all the included samples.

In contrast, a further embodiment of this invention is the computerized form of this process, termed here "electronic northern blot." In this variation, a single gene is queried for expression against a multitude of prepared and sequenced libraries present within the database. In this way, the pattern of expression of any single candidate gene can be examined instantaneously and effortlessly. More candidate genes can thus be scanned, leading to more frequent and fruitfully relevant discoveries. The computer program included as Table 5 includes a program for performing this function, and Table 6 is a partial listing of entries of the database used in the electronic northern blot analysis.

6.11. PHASE I CLINICAL TRIALS

Based on the establishment of safety and effectiveness in the above animal tests, Phase I clinical tests are undertaken. Normal patients are subjected to the usual preliminary clinical laboratory tests. In addition, appropriate specimens are taken and subjected to gene transcript analysis. Additional patient specimens are taken at predetermined intervals during the test. The specimens are subjected to gene transcript analysis as described above. In addition, the gene transcript changes noted in the earlier rat toxicity study are carefully evaluated as clinical markers in the followed patients. Changes in the gene transcript analyses are evaluated as indicators of toxicity by correlation with clinical signs and symptoms and other laboratory results. In addition, subtraction is performed on individual patient specimens and on averaged patient specimens. The subtraction analysis highlights any toxicological changes in the treated patients. This is a highly refined determinant of toxicity. The subtraction method also annotates clinical markers. Further subgroups can be analyzed by subtraction analysis, including, for example, 1) segregation by

occurrence and type of adverse effect; and 2) segregation by dosage.

6.12. GENE TRANSCRIPT IMAGING ANALYSIS IN CLINICAL STUDIES

A gene transcript imaging analysis (or multiple gene transcript imaging analyses) is a useful tool in other clinical studies. For example, the differences in gene transcript imaging analyses before and after treatment can be assessed for patients on placebo and drug treatment. This method also effectively screens for clinical markers to follow in clinical use of the drug.

6.13. COMPARATIVE GENE TRANSCRIPT ANALYSIS BETWEEN SPECIES

The subtraction method can be used to screen cDNA libraries from diverse sources. For example, the same cell types from different species can be compared by gene transcript analysis to screen for specific differences, such as in detoxification enzyme systems. Such testing aids in the selection and validation of an animal model for the commercial purpose of drug screening or toxicological testing of drugs intended for human or animal use. When the comparison between animals of different species is shown in columns for each species, we refer to this as an interspecies comparison, or zoo blot.

Embodiments of this invention may employ databases such as those written using the FoxBASE programming language commercially available from Microsoft Corporation. Other embodiments of the invention employ other databases, such as a random peptide database, a polymer database, a synthetic oligomer database, or a oligonucleotide database of the type described in U.S. Patent 5,270,170, issued December 14, 1993 to Cull, et al., PCT International Application Publication No. WO 9322684, published November 11, 1993, PCT International Application Publication No. WO 9306121, published April 1, 1993, or PCT International Application Publication No. WO 9119818, published December 26, 1991. These four references (whose text is incorporated herein by reference) include teaching which

may be applied in implementing such other embodiments of the present invention.

All references referred to in the preceding text are hereby expressly incorporated by reference herein.

Various modifications and variations of the described method and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments.

TABLE 1

Designations (D)

E = Exact
H = Homologous
O = Other species
N = No match
D = Noncoding gene
U = Nonreadable
R = Repetitive DNA
A = Poly-A only
V = Vector only
M = Mitochondrial DNA
S = Skip
I = Match Incyte clone
X = EST match

Library (L)

U = U937
M = HMC
T = THP-1
H = HUVEC
S = Spleen
L = Lung
Y = T & B cell
A = Adenoid

Distribution (F)

C = Non-specific
P = Cell/tissue specific
U = Unknown

Species (S)

H = Human
A = Ape
P = Pig
D = Dog
V = Bovine
B = Rabbit
R = Rat
M = Mouse
S = Hamster
C = Chicken
F = Amphibian
I = Invertebrate
Z = Protozoan
G = Fungi

Localization (Z)

N = Nuclear
C = Cytoplasmic
K = Cytoskeleton
E = Cell surface
Z = Intracellular memb
M = Mitochondrial
S = Secreted
U = Unknown
X = Other

Status (I)

0 = No current interest
1 = Do primary analysis
2 = Primary analysis done
3 = Full length sequence
4 = Secondary analysis
5 = Tissue northern
6 = Obtain full length

Function (R)

T = Translation
L = Protein processing
R = Ribosomal protein
O = *Oncogene*
G = GTP binding ptn
V = Viral element
Y = Kinase/phosphatase
A = Tumor antigen related
I = *Binding proteins*
D = NA-binding /transcription
B = Surface molecule/receptor
C = Ca^{++} binding protein
S = Ligands/effectors
H = Stress response protein
E = *Enzyme*
F = Ferroprotein
P = Protease/inhibitor
Z = Oxidative phosphorylation
Q = Sugar metabolism
M = Amino acid metabolism
N = Nucleic acid metabolism
W = Lipid metabolism
K = Structural
X = Other
U = unknown

TABLE 2

Clone numbers 15000 through 20000
Libraries: HUVEC
Arranged by ABUNDANCE
Total clones analyzed: 5000

319 genes, for a total of 1713 Clones

	number	N	c	entry	s	descriptor
1	15365	67		HSRPL41		Riboptn L41
2	15004	65		NCY015004		INCYTE 015004
3	15638	63		NCY015638		INCYTE 015638
4	15390	50		NCY015390		INCYTE 015390
5	15193	47		HSFIB1		Fibronectin
6	15220	47		RRRPL9	R	Riboptn L9
7	15280	47		NCY015280		INCYTE 015280
8	15583	33		M62060		EST HHCH09 (IGR)
9	15662	31		HSACTCGR		Actin, gamma
10	15026	29		NCY015026		INCYTE 015026
11	15279	24		HSEF1AR		Elf 1-alpha
12	15027	23		NCY015027		INCYTE 015027
13	15033	20		NCY015033		INCYTE 015033
14	15198	20		NCY015198		INCYTE 015198
15	15809	20		HSCOLL1		Collagenase
16	15221	19		NCY015221		INCYTE 015221
17	15263	19		NCY015263		INCYTE 015263
18	15290	19		NCY015290		INCYTE 015290
19	15350	18		NCY015350		INCYTE 015350
20	15030	17		NCY015030		INCYTE 015030
21	15234	17		NCY015234		INCYTE 015234
22	15459	16		NCY015459		INCYTE 015459
23	15353	15		NCY015353		INCYTE 015353
24	15378	15		S76965		Ptn kinase inhib
25	15255	14		HUMTHYB4		Thymosin beta-4
26	15401	14		HSLIPCR		Lipocortin I
27	15425	14		HSPOLYAB		Poly-A bp
28	18212	14		HUMTHYMA		Thymosin, alpha
29	18216	14		HSMRP1		Motility relat ptn; MRP-1;CD-9
30	15189	13		HS18D		Interferon induc ptn 1-8D
31	15031	12		HUMFKBP		FK506 bp
32	15306	12		HSH2AZ		Histone H2A
33	15621	12		HUMLEC		Lectin, B-galbp, 14kDa
34	15789	11		NCY015789		INCYTE 015789
35	16578	11		HSRPS11		Riboptn S11
36	16632	11		M61984		EST HHCA13 (IGR)
37	18314	11		NCY018314		INCYTE 018314
38	15367	10		NCY015367		INCYTE 015367
39	15415	10		HSIFNIN1		interferon induc mRNA
40	15633	10		HSLDHAR		Lactate dehydrogenase
41	15813	10		CHKNMHCB		C Myosin heavy chain B
42	18210	10		NCY018210		INCYTE 018210
43	18233	10		HSRPII140		RNA polymerase II
44	18996	10		NCY018996		INCYTE 018996
45	15088	9		HUMFERL		Ferritin, light chain
46	15714	9		NCY015714		INCYTE 015714
47	15720	9		NCY015720		INCYTE 015720
48	15863	9		NCY015863		INCYTE 015863
49	16121	9		HSET		Endothelin
50	18252	9		NCY018252		INCYTE 018252
51	15351	8		HUMALBP		Lipid bp, adipocyte
52	15370	8		NCY015370		INCYTE 015370

TABLE 2 Con't

	number	N	c	entry	s	descriptor
53	15670	8		BTCIASHI	V	NADH-ubiq oxidoreductase
54	15795	8		NCY015795		INCYTE 015795
55	16245	8		NCY016245		INCYTE 016245
56	18262	8		NCY018262		INCYTE 018262
57	18321	8		HSRPL17		Riboptn L17
58	15126	7		XLRPL1BRF		Riboptn L1
59	15133	7		HSAC07		Actin, beta
60	15245	7		NCY015245		INCYTE 015245
61	15288	7		NCY015288		INCYTE 015288
62	15294	7		HSGAPDR		G-3-PD
63	15442	7		HUMLAMB		Laminin receptor, 54kDa
64	15485	7		HSNGMRNA		Uracil DNA glycosylase
65	16646	7		NCY016646		INCYTE 016646
66	18003	7		HUMPAIA		Plsmnogen activ gene
67	15032	6		HUMUB		Ubiquitin
68	15267	6		HSRPS8		Riboptn S8
69	15295	6		NCY015295		INCYTE 015295
70	15458	6		RNRPS10R	R	Riboptn S10
71	15832	6		RSGALEM	R	UDP-galactose epimerase
72	15928	6		HUMAPOJ		Apolipoptn J
73	16598	6		HUMTBBM40		Tubulin, beta
74	18218	6		NCY018218		INCYTE 018218
75	18499	6		HSP27		Hydrophobic ptn p27
76	18963	6		NCY018963		INCYTE 018963
77	18997	6		NCY018997		INCYTE 018997
78	15432	5		HSAGALAR		Galactosidase A, alpha
79	15475	5		NCY015475		INCYTE 015475
80	15721	5		NCY015721		INCYTE 015721
81	15865	5		NCY015865		INCYTE 015865
82	16270	5		NCY016270		INCYTE 016270
83	16886	5		NCY016886		INCYTE 016886
84	18500	5		NCY018500		INCYTE 018500
85	18503	5		NCY018503		INCYTE 018503
86	19672	5		RRRPL34	R	Riboptn L34
87	15086	4		XLRPL1AR	F	Riboptn L1a
88	15113	4		HUMIFNWRS		tRNA synthetase, trp
89	15242	4		NCY015242		INCYTE 015242
90	15249	4		NCY015249		INCYTE 015249
91	15377	4		NCY015377		INCYTE 015377
92	15407	4		NCY015407		INCYTE 015407
93	15473	4		NCY015473		INCYTE 015473
94	15588	4		HSRPS12		Riboptn S12
95	15684	4		HSEF1G		Elf 1-gamma
96	15782	4		NCY015782		INCYTE 015782
97	15916	4		HSRPS18		Riboptn S18
98	15930	4		NCY015930		INCYTE 015930
99	16108	4		NCY016108		INCYTE 016108
100	16133	4		NCY016133		INCYTE 016133

NORMAL MONONCYTE VS. ACTIVATED MACROPHAGE

Top 15 Most Abundant Genes

	NORMAL	ACTIVATED
1	Elongation factor-1 alpha	Interleukin-1 beta
2	Ribosomal phosphoprotein	Macrophage inflammatory protein-1
3	Ribosomal protein S8 homolog	Interleukin-8
4	Beta-Globin	Lymphocyte activation gene
5	Ferritin H chain	Elongation factor-1 alpha
6	Ribosomal protein L7	Beta actin
7	Nucleoplasmin	Rantes T-cell specific protein
8	Ribosomal protein S20 homolog	Poly A binding protein
9	Transferrin receptor	Osteopontin; nephropontin
10	Poly-A binding protein	Tumor Necrosis Factor-alpha
11	Translationally controlled tumor ptn	INCYTE clone 011050
12	Ribosomal protein S25	Cu/Zn superoxide dismutase
13	Signal recognition particle SRP9	Adenylate cyclase (yeast homolog)
14	Histone H2A.Z	NGF-related B cell activation molecule
15	Ribosomal protein Ke-3	Protease Nexin-1, glial-derived

TABLE 3

TABLE 4

Libraries: THP-1
Subtracting: HMC
Sorted by ABUNDANCE
Total clones analyzed: 7375

1057 genes, for a total of 2151 clones

number	entry	s	descriptor	bgfreq	rfend	ratio
10022	HUMIL1		IL 1-beta	0	131	262.00
10036	HSMDNCF		IL-8	0	119	238.00
10089	HSLAG1CDN		Lymphocyte activ gene	0	71	142.00
10060	HUMTCSM		RANTES	0	23	46.000
10003	HUMMIP1A		MIP-1	3	121	40.333
10689	HSOP		Osteopontin	0	20	40.000
11050	NCY011050		INCYTE 011050	0	17	34.000
10937	HSTNFR		TNF-alpha	0	17	34.000
10176	HSSOD		Superoxide dismutase	0	14	28.000
10886	HSCDW40		B-cell activ,NGF-relat	0	10	20.000
10186	HUMAPR		Early resp PMA-induc	0	9	18.000
10967	HUMGDN		PN-1, glial-deriv	0	9	18.000
11353	NCY011353		INCYTE 011353	0	8	16.000
10298	NCY010298		INCYTE 010298	0	7	14.000
10215	HUM4COLA		Collagenase, type IV	0	6	12.000
10276	NCY010276		INCYTE 010276	0	6	12.000
10488	NCY010488		INCYTE 010488	0	6	12.000
11138	NCY011138		INCYTE 011138	0	6	12.000
10037	HUMCAPPRO		Adenylate cyclase	1	10	10.000
10840	HUMADCY		Adenylate cyclase	0	5	10.000
10672	HSCD44E		Cell adhesion glptn	0	5	10.000
12837	HUMCYCLOX		Cyclooxygenase-2	0	5	10.000
10001	NCY010001		INCYTE 010001	0	5	10.000
10005	NCY010005		INCYTE 010005	0	5	10.000
10294	NCY010294		INCYTE 010294	0	5	10.000
10297	NCY010297		INCYTE 010297	0	5	10.000
10403	NCY010403		INCYTE 010403	0	5	10.000
10699	NCY010699		INCYTE 010699	0	5	10.000
10966	NCY010966		INCYTE 010966	0	5	10.000
12092	NCY012092		INCYTE 012092	0	5	10.000
12549	HSRHOB		Oncogene rho	0	5	10.000
10691	HUMARF1BA		ADP-ribosylation fctr	0	4	8.000
12106	HSADSS		Adenylosuccinate synthetase	0	4	8.000
10194	HSCATHL		Cathepsin L	0	4	8.000
10479	CLMCYCA	I	Cyclin A	0	4	8.000
10031	NCY010031		INCYTE 010031	0	4	8.000
10203	NCY010203		INCYTE 010203	0	4	8.000
10288	NCY010288		INCYTE 010288	0	4	8.000
10372	NCY010372		INCYTE 010372	0	4	8.000
10471	NCY010471		INCYTE 010471	0	4	8.000
10484	NCY010484		INCYTE 010484	0	4	8.000
10859	NCY010859		INCYTE 010859	0	4	8.000
10890	NCY010890		INCYTE 010890	0	4	8.000
11511	NCY011511		INCYTE 011511	0	4	8.000
11868	NCY011868		INCYTE 011868	0	4	8.000
12820	NCY012820		INCYTE 012820	0	4	8.000
10133	HSI1RAP		IL-1 antagonist	0	4	8.000
10516	HUMP2A		Phosphatase, regul 2A	0	4	8.000
11063	HUMB94		TNF-induc response	0	4	8.000
11140	HSHB15RNA		HB15 gene; new Ig	0	3	6.000
10788	NCY001713		INCYTE 001713	0	3	6.000
10033	NCY010033		INCYTE 010033	0	3	6.000
10035	NCY010035		INCYTE 010035	0	3	6.000
10084	NCY010084		INCYTE 010084	0	3	6.000
10236	NCY010236		INCYTE 010236	0	3	6.000
10383	NCY010383		INCYTE 010383	0	3	6.000

TABLE 4 Con't

number	entry	s	descriptor	bgfreq	rfend	ratio
10450	NCY010450		INCYTE 010450	0	3	6.000
10470	NCY010470		INCYTE 010470	0	3	6.000
10504	NCY010504		INCYTE 010504	0	3	6.000
10507	NCY010507		INCYTE 010507	0	3	6.000
10598	NCY010598		INCYTE 010598	0	3	6.000
10779	NCY010779		INCYTE 010779	0	3	6.000
10909	NCY010909		INCYTE 010909	0	3	6.000
10976	NCY010976		INCYTE 010976	0	3	6.000
10985	NCY010985		INCYTE 010985	0	3	6.000
11052	NCY011052		INCYTE 011052	0	3	6.000
11068	NCY011068		INCYTE 011068	0	3	6.000
11134	NCY011134		INCYTE 011134	0	3	6.000
11136	NCY011136		INCYTE 011136	0	3	6.000
11191	NCY011191		INCYTE 011191	0	3	6.000
11219	NCY011219		INCYTE 011219	0	3	6.000
11386	NCY011386		INCYTE 011386	0	3	6.000
11403	NCY011403		INCYTE 011403	0	3	6.000
11460	NCY011460		INCYTE 011460	0	3	6.000
11618	NCY011618		INCYTE 011618	0	3	6.000
11686	NCY011686		INCYTE 011686	0	3	6.000
12021	NCY012021		INCYTE 012021	0	3	6.000
12025	NCY012025		INCYTE 012025	0	3	6.000
12320	NCY012320		INCYTE 012320	0	3	6.000
12330	NCY012330		INCYTE 012330	0	3	6.000
12853	NCY012853		INCYTE 012853	0	3	6.000
14386	NCY014386		INCYTE 014386	0	3	6.000
14391	NCY014391		INCYTE 014391	0	3	6.000

TABLE 5

```
* Master menu for SUBTRACTION output
SET TALK OFF
SET SAFETY OFF
SET EXACT ON
SET TYPEAHEAD TO 0
CLEAR
SET DEVICE TO SCREEN
USE "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
GO TOP
STORE NUMBER TO INITIATE
GO BOTTOM
STORE NUMBER TO TERMINATE
STORE '          ' TO Target1
STORE '          ' TO Target2
STORE '          ' TO Target3
STORE '          ' TO Object1
STORE '          ' TO Object2
STORE '          ' TO Object3
STORE 0 TO ANAL
STORE 0 TO EMATCH
STORE 0 TO HMATCH
STORE 0 TO OMATCH
STORE 0 TO IMATCH
STORE 0 TO PTF
STORE 1 TO BAIL
DO WHILE .T.
* Program...: Subtraction 2.fmt
* Date....: 10/11/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes....: Format file Subtraction 2
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",9 COLOR 0,0,0,
@ PIXELS 75,120 TO 178,241 STYLE 3871 COLOR 0,0,-1,24610,-1,8947
@ PIXELS 27,134 SAY "Subtraction Menu" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 117,126 GET EMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Exact " SIZE 15,62 CO
@ PIXELS 135,126 GET HMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Homologous" SIZE 15,1
@ PIXELS 153,126 GET OMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Other spc" SIZE 15,84
@ PIXELS 90,152 SAY "Matches:" STYLE 65535 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 171,126 GET Imatch STYLE 65536 FONT "Chicago",12 PICTURE "@*C Incyte" SIZE 15,65 CO
@ PIXELS 252,137 GET initiate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 252,236 GET terminate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 252,35 SAY "Include clones " STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 252,215 SAY "->" STYLE 65536 FONT "Geneva",14 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 198,126 GET PTF STYLE 65536 FONT "Chicago",12 PICTURE "@*C Print to file" SIZE 15,9
@ PIXELS 90,9 TO 181,109 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 90,288 TO 181,397 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 81,296 SAY "Background:" STYLE 65536 FONT "Geneva",270 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 45,135 GET ANAL STYLE 65536 FONT "Chicago",12 PICTURE "@*R Overall;Function" SIZE 4
@ PIXELS 81,26 SAY "Target:" STYLE 65536 FONT "Geneva",270 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 108,20 GET target1 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 135,20 GET target2 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 162,20 GET target3 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 108,299 GET object1 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 135,299 GET object2 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 162,299 GET object3 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 276,324 GET Bail STYLE 65536 FONT "Chicago",12 PICTURE "@*R Run;Bail out" SIZE 4112
*
* EOF: Subtraction 2.fmt
READ
  IF Bail=2
  CLEAR
  CLOSE DATABASES
  USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
  SET SAFETY ON
  SCREEN 1 OFF
  RETURN
```

```
   ENDIF
 STORE VAL(SYS(2)) TO STARTIME
 STORE UPPER(Target1) TO Target1
 STORE UPPER(Target2) TO Target2
 STORE UPPER(Target3) TO Target3
 STORE UPPER(Object1) TO Object1
 STORE UPPER(Object2) TO Object2
 STORE UPPER(Object3) TO Object3
 clear
 SET TALK ON
 GAP = TERMINATE-INITIATE+1
 GO INITIATE
 COPY NEXT GAP FIELDS NUMBER,library,D,F,Z,R,ENTRY,S,DESCRIPTOR,START,RFEND,I TO TEMPNUM
 USE TEMPNUM
 COUNT TO TOT
 COPY TO TEMPRED FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='I'
 USE TEMPRED

 IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
 COPY TO TEMPDESIG
 ELSE
 COPY STRUCTURE TO TEMPDESIG
 USE TEMPDESIG
   IF Ematch=1
   APPEND FROM TEMPNUM FOR D='E'
   ENDIF
   IF Hmatch=1
   APPEND FROM TEMPNUM FOR D='H'
   ENDIF
   IF Omatch=1
   APPEND FROM TEMPNUM FOR D='O'
   ENDIF
   IF Imatch=1
   APPEND FROM TEMPNUM FOR D='I'.OR.D='X'
*.OR.D='N'
  ENDIF
ENDIF
COUNT TO STARTOT

COPY STRUCTURE TO TEMPLIB
USE TEMPLIB
  APPEND FROM TEMPDESIG FOR library=UPPER(target1)
  IF target2<>'          '
  APPEND FROM TEMPDESIG FOR library=UPPER(target2)
  ENDIF
  IF target3<>'          '
  APPEND FROM TEMPDESIG FOR library=UPPER(target3)
  ENDIF
COUNT TO ANALTOT

USE TEMPDESIG
COPY STRUCTURE TO TEMPSUB
USE TEMPSUB
  APPEND FROM TEMPDESIG FOR library=UPPER(Object1)
  IF target2<>'          '
  APPEND FROM TEMPDESIG FOR library=UPPER(Object2)
  ENDIF
  IF target3<>'
  APPEND FROM TEMPDESIG FOR library=UPPER(Object3)
  ENDIF
COUNT TO SUBTRACTOT
SET TALK OFF
*****************************************************************************************
* COMPRESSION SUBROUTINE A
? 'COMPRESSING QUERY LIBRARY'
USE TEMPLIB
```

```
SORT ON ENTRY,NUMBER TO LIBSORT
USE LIBSORT
COUNT TO IDGENE
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= IDGENE
  PACK
  COUNT TO AUNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
STORE D TO DESIGA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
STORE D TO DESIGB
  IF TESTA = TESTB.AND.DESIGA=DESIGB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
SORT ON RFEND/D,NUMBER TO TEMPTARSORT
USE TEMPTARSORT
*REPLACE ALL START WITH RFEND/IDGENE*10000
COUNT TO TEMPTARCO
*************************************************************************************
* COMPRESSION SUBROUTINE B
? 'COMPRESSING TARGET LIBRARY'
USE TEMPSUB
SORT ON ENTRY,NUMBER TO SUBSORT
USE SUBSORT
COUNT TO SUBGENE
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= SUBGENE
  PACK
  COUNT TO BUNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
STORE D TO DESIGA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
STORE D TO DESIGB
  IF TESTA = TESTB.AND.DESIGA=DESIGB
```

```
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
 GO MARK1
 REPLACE RFEND WITH DUP
 MARK1 = MARK1+DUP
 SW=1
 LOOP
 ENDDO TEST
 LOOP
 ENDDO ROLL
 SORT ON RFEND/D,NUMBER TO TEMPSUBSORT
 USE TEMPSUBSORT
 *REPLACE ALL START WITH RFEND/IDGENE*10000
 COUNT TO TEMPSUBCO
 ************************************************************************************************
 *FUSION ROUTINE
 ? 'SUBTRACTING LIBRARIES'
 USE SUBTRACTION
 COPY STRUCTURE TO CRUNCHER
 SELECT 2
 USE TEMPSUBSORT
 SELECT 1
 USE CRUNCHER
 APPEND FROM TEMPTARSORT
 COUNT TO BAILOUT
 MARK = 0

 DO WHILE .T.
 SELECT 1
 MARK = MARK+1
   IF MARK>BAILOUT
   EXIT
   ENDIF
 GO MARK
 STORE ENTRY TO SCANNER
 SELECT 2
 LOCATE FOR ENTRY=SCANNER
 IF FOUND()
 STORE RFEND TO BIT1
 STORE RFEND TO BIT2
 ELSE
 STORE 1/2 TO BIT1
 STORE 0 TO BIT2
 ENDIF
 SELECT 1
 REPLACE BGFREQ WITH BIT2
 REPLACE ACTUAL WITH BIT1
 LOOP
 ENDDO

 SELECT 1
 REPLACE ALL RATIO WITH RFEND/ACTUAL
 ? 'DOING FINAL SORT BY RATIO'
 SORT ON RATIO/D,BGFREQ/D,DESCRIPTOR TO FINAL
 USE FINAL
 ************************************************************************************************
 set talk off
 DO CASE
 CASE PIF=0
 SET DEVICE TO PRINT
 SET PRINT ON
 EJECT
 CASE PIF=1
 SET ALTERNATE TO 'Adenoid Patent Figures:Subtraction.txt'
```

```
SET ALTERNATE ON
ENDCASE

STORE VAL(SYS(2)) TO FINTIME
IF FINTIME<STARTIME
STORE FINTIME+86400 TO FINTIME
ENDIF
STORE FINTIME - STARTIME TO COMPSEC
STORE COMPSEC/60 TO COMPMIN

**********************
SET MARGIN TO 10
@1,1 SAY "Library Subtraction Analysis" STYLE 65536 FONT "Geneva",274 COLOR 0,0,0,-1,-1,-1
?
?
?
?
? date()
?? '  '
?? TIME()
? 'Clone numbers '
?? STR(INITIATE,5,0)
?? ' through '
?? STR(TERMINATE,6,0)
? 'Libraries: '
? Target1
IF Target2<>'
?? ','
?? Target2
ENDIF
IF Target3<>'
?? ', '
?? Target3
ENDIF
? 'Subtracting:
? Object1
IF Object2<>'
?? ','
?? Object2
ENDIF
IF Object3<>'
?? ', '
?? Object3
ENDIF
? 'Designations: '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other sp.'
ENDIF
IF Imatch=1
?? 'INCYTE'
ENDIF
IF ANAL=1
? 'Sorted by ABUNDANCE'
ENDIF
IF ANAL=2
? 'Arranged by FUNCTION'
ENDIF
```

```
? 'Total clones represented: '
?? STR(TOT,5,0)
? 'Total clones analyzed: '
?? STR(STARTOT,5,0)
? 'Total computation time:
?? STR(COMPMIN,5,2)
?? ' minutes'
?
? 'd = designation   f = distribution   z = location   r = function   s = species   i = inte
?
**************************
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',9 COLOR 0,0,0,
DO CASE
CASE ANAL=1
?? STR(AUNIQUE,4,0)
?? ' genes, for a total of '
?? STR(ANALTOT,4,0)
?? ' clones'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I
SET PRINT OFF
CLOSE DATABASES
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'

CASE ANAL=2
* arrange/function
SET PRINT ON
SET HEADING ON
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',268 COLOR 0
?
? '                              BINDING PROTEINS'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Surface molecules and receptors:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='B'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Calcium-binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='C'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Ligands and effectors:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='S'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Other binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='I'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',268 COLOR 0
? '                              ONCOGENES'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'General oncogenes:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='O'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'GTP-binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='G'
```

```
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Viral elements:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='V'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Kinases and Phosphatases:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='Y'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Tumor-related antigens:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='A'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',268 COLOR 0
? '          PROTEIN SYNTHETIC MACHINERY PROTEINS'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Transcription and Nucleic Acid-binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='D'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Translation:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='T'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Ribosomal proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='R'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Protein processing:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='L'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',268 COLOR 0
?
?                              ENZYMES'
?

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Ferroproteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='F'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Proteases and inhibitors:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='P'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Oxidative phosphorylation:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='Z'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Sugar metabolism:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='Q'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Amino acid metabolism:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
```

```
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='M'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Nucleic acid metabolism:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='N'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Lipid metabolism:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='W'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Other enzymes:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='E'
?
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",268 COLOR 0
?
? '                          MISCELLANEOUS CATEGORIES'
?

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Stress response:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='H'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Structural:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='K'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Other clones:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='X'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Clones of unknown function:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='U'
ENDCASE

DO "Test print.prg"
SET PRINT OFF
SET DEVICE TO SCREEN
CLOSE DATABASES
ERASE TEMPLIB.DBF
ERASE TEMPNUM.DBF
ERASE TEMPDESIG.DBF
SET MARGIN TO 0
CLEAR
LOOP
ENDDO
```

```
*Northern (single), version 11-25-94
close databases
SET TALK OFF
SET PRINT OFF
SET EXACT OFF
CLEAR
STORE '           ' TO Eobject
STORE '                                    ' TO Dobject
STORE 0 TO Numb
STORE 0 TO Zog
STORE 1 TO Bail
DO WHILE .T.
* Program.: Northern (single).fmt
* Date....:  8/ 8/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes...: Format file Northern (single)
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",12 COLOR 0,0,0
@ PIXELS 15,81 TO 46,397 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 89,79 TO 192,422 STYLE 28447 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 115,98 SAY "Entry #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 115,173 GET Eobject STYLE 0 FONT "Geneva",12 SIZE 15,142 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,89 SAY "Description" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,173 GET Dobject STYLE 0 FONT "Geneva",12 SIZE 15,241 COLOR 0,0,0,-1,-1,-1
@ PIXELS 35,89 SAY "Single Northern search screen" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-
@ PIXELS 220,162 GET Bail STYLE 65536 FONT "Chicago",12 PICTURE "@*R Continue;Bail out" SIZE
@ PIXELS 175,98 SAY "Clone #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 175,173 GET Numb STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,0,-1,-1,-1
@ PIXELS 80,152 SAY "Enter any ONE of the following:" STYLE 65536 FONT "Geneva",12 COLOR -1,
*
* EOP: Northern (single).fmt
READ
IF Bail=2
CLEAR
screen 1 off
RETURN
ENDIF
USE "SmartGuy:FoxBASE+/Mac:Fox files:Lookup.dbf"
SET TALK ON

IF Eobject<>'
STORE UPPER(Eobject) to Eobject
SET SAFETY OFF
SORT ON Entry TO "Lookup entry.dbf"
SET SAFETY ON
USE "Lookup entry.dbf"
LOCATE FOR Look=Eobject
IF .NOT.FOUND()
CLEAR
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup entry.dbf"
ENDIF

IF Dobject<>'
SET EXACT OFF
SET SAFETY OFF
SORT ON descriptor TO "Lookup descriptor.dbf"
SET SAFETY On
USE "Lookup descriptor.dbf"
LOCATE FOR UPPER(TRIM(descriptor))=UPPER(TRIM(Dobject))
IF .NOT.FOUND()
CLEAR
```

```
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE 'Lookup descriptor.dbf'
SET EXACT ON
ENDIF

IF Numb<>0
USE 'SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf'
GO Numb
BROWSE
STORE Entry TO Searchval
ENDIF

CLEAR
? 'Northern analysis for entry '
?? Searchval
?
? 'Enter Y to proceed'
WAIT TO OK
CLEAR
IF UPPER(OK)<>'Y'
screen 1 off
RETURN
ENDIF

* COMPRESSION SUBROUTINE FOR Library.dbf
? 'Compressing the Libraries file now...'
USE 'SmartGuy:FoxBASE+/Mac:Fox files:libraries.dbf'
SET SAFETY OFF
SORT ON library TO 'Compressed libraries.dbf'
* FOR entered>0
SET SAFETY ON
USE 'Compressed libraries.dbf'
DELETE FOR entered=0
PACK
COUNT TO TOT
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
   SW2=1
   LOOP
   ENDIF
GO MARK1
STORE library TO TESTA
SKIP
STORE Library TO TESTB
IF TESTA = TESTB
DELETE
ENDIF
MARK1 = MARK1+1
LOOP
ENDDO ROLL

* Northern analysis
CLEAR
? 'Doing the northern now...
SET TALK ON
USE 'SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf'
SET SAFETY OFF
COPY TO 'Hits.dbf' FOR entry=searchval
SET SAFETY ON
```

```
* MASTER ANALYSIS 3; VERSION 12-9-94
* Master menu for analysis output
CLOSE DATABASES
SET TALK OFF
SET SAFETY OFF
CLEAR
SET DEVICE TO SCREEN
SET DEFAULT TO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:"
USE "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
GO TOP
STORE NUMBER TO INITIATE
GO BOTTOM
STORE NUMBER TO TERMINATE
STORE 0 TO ENTIRE
STORE 0 TO CONDEN
STORE 0 TO ANAL
STORE 0 TO EMATCH
STORE 0 TO HMATCH
STORE 0 TO OMATCH
STORE 0 TO IMATCH
STORE 0 TO XMATCH
STORE 0 TO PRINTON
STORE 0 TO PTF
DO WHILE .T.
* Program.: Master analysis.fmt
* Date....: 12/ 9/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes....: Format file Master analysis
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",9 COLOR 0,0,0,
@ PIXELS 39,255 TO 277,430 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 75,120 TO 178,241 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 27,98 SAY "Customized Output Menu" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-1,-1,-1
@ PIXELS 45,54 GET conden STYLE 65536 FONT "Chicago",12 PICTURE "@*C Condensed format" SIZE
@ PIXELS 54,261 GET anal STYLE 65536 FONT "Chicago",12 PICTURE "@*RV Sort/number;Sort/entry;
@ PIXELS 117,126 GET EMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Exact " SIZE 15,62 CO
@ PIXELS 135,126 GET HMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Homologous" SIZE 15,1
@ PIXELS 153,126 GET OMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Other spc" SIZE 15,84
@ PIXELS 90,152 SAY "Matches:" STYLE 65536 FONT "Geneva",268 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 63,54 GET PRINTON STYLE 65536 FONT "Chicago",12 PICTURE "@*C Include clone listing"
@ PIXELS 171,126 GET Imatch STYLE 65536 FONT "Chicago",12 PICTURE "@*C Incyte" SIZE 15,65 CO
@ PIXELS 252,146 GET initiate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 270,146 GET terminate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 234,134 SAY "Include clones " STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 270,125 SAY "->" STYLE 65536 FONT "Geneva",14 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 198,126 GET PTF STYLE 65536 FONT "Chicago",12 PICTURE "@*C Print to file" SIZE 15,9
@ PIXELS 189,0 TO 257,120 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 209,8 SAY "Library selection" STYLE 65536 FONT "Geneva",266 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 227,18 GET ENTIRE STYLE 65536 FONT "Chicago",12 PICTURE "@*RV All;Selected" SIZE 16
*
* EOF: Master analysis.fmt
READ
  IF ANAL=9
  CLEAR
  CLOSE DATABASES
  ERASE TEMPMASTER.DBF
  USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
  SET SAFETY ON
  SCREEN 1 OFF
  RETURN
  ENDIF
clear
? INITIATE
? TERMINATE
? CONDEN
? ANAL
```

```
? ematch
? Hmatch
? Omatch
? IMATCH
SET TALK ON
  IF ENTIRE=2
USE "Unique libraries.dbf"
  REPLACE ALL i WITH ' '
  BROWSE FIELDS i,libname,library,total,entered AT 0,0
  ENDIF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
*COPY TO TEMPNUM FOR NUMBER>=INITIATE.AND.NUMBER<=TERMINATE
*USE TEMPNUM
COPY STRUCTURE TO TEMPLIB
USE TEMPLIB
  IF ENTIRE=1
  APPEND FROM "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
  ENDIF
  IF ENTIRE=2
USE "Unique libraries.dbf"
  COPY TO SELECTED FOR UPPER(i)='Y'
  USE SELECTED
  STORE RECCOUNT() TO STOPIT
  MARK=1
    DO WHILE .T.
    IF MARK>STOPIT
    CLEAR
    EXIT
    ENDIF
    USE SELECTED
    GO MARK
    STORE library TO THISONE
    ? 'COPYING '
    ?? THISONE
    USE TEMPLIB
    APPEND FROM "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf" FOR library=THISONE
    STORE MARK+1 TO MARK
    LOOP
    ENDDO
  ENDIF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
COUNT TO STARTOT
COPY STRUCTURE TO TEMPDESIG
USE TEMPDESIG
  IF Ematch=0 .AND.. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
  APPEND FROM TEMPLIB
  ENDIF
  IF Ematch=1
  APPEND FROM TEMPLIB FOR D='E'
  ENDIF
  IF Hmatch=1
  APPEND FROM TEMPLIB FOR D='H'
  ENDIF
  IF Omatch=1
  APPEND FROM TEMPLIB FOR D='O'
  ENDIF
  IF Imatch=1
  APPEND FROM TEMPLIB FOR D='I'.OR.D='X'.OR.D='N'
  ENDIF
  IF Xmatch=1
  APPEND FROM TEMPLIB FOR D='X'
  ENDIF
COUNT TO ANALTOT
set talk off
**********************
DO CASE
```

```
CASE PTF=0
SET DEVICE TO PRINT
SET PRINT ON
EJECT
CASE PTF=1
SET ALTERNATE TO "Total function sort.txt"
*SET ALTERNATE TO "H and O function sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Abundance sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Abundance con.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Function sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Distribution sort.txt"
*SET ALTERNATE TO "Shear stress HUVEC 1:Clone list.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Location sort.txt"
SET ALTERNATE ON
ENDCASE
**********************
IF PRINTON=1
@1,30 SAY "Database Subset Analysis" STYLE 65536 FONT "Geneva",274 COLOR 0,0,0,-1,-1,-1
ENDIF
?
?
?
?
? date()
?? '    '
?? TIME()
? 'Clone numbers '
?? STR(INITIATE,6,0)
?? ' through '
?? STR(TERMINATE,6,0)
? 'Libraries: '
IF ENTIRE=1
? 'All libraries'
ENDIF
IF ENTIRE=2
    MARK=1
    DO WHILE .T.
    IF MARK>STOPIT
    EXIT
    ENDIF
    USE SELECTED
    GO MARK
    ? '  '
    ?? TRIM(libname)
    STORE MARK+1 TO MARK
    LOOP
    ENDDO
ENDIF
? 'Designations: '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other sp.'
ENDIF
IF Imatch=1
?? 'INCYTE'
ENDIF
IF Xmatch=1
?? 'EST'
```

```
ENDIF
IF CONDEN=1
? 'Condensed format analysis'
ENDIF
IF ANAL=1
? 'Sorted by NUMBER'
ENDIF
IF ANAL=2
? 'Sorted by ENTRY'
ENDIF
IF ANAL=3
? 'Arranged by ABUNDANCE'
ENDIF
IF ANAL=4
? 'Sorted by INTEREST'
ENDIF
IF ANAL=5
? 'Arranged by LOCATION'
ENDIF
IF ANAL=6
? 'Arranged by DISTRIBUTION'
ENDIF
IF ANAL=7
? 'Arranged by FUNCTION'
ENDIF
? 'Total clones represented: '
?? STR(STARTOT,6,0)
? 'Total clones analyzed: '
?? STR(ANALTOT,6,0)
?
? 'l = library    d = designation   f = distribution   z = location   r = function   c = cer
?
*********************************************************************************************
USE TEMPDESIG
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
DO CASE
CASE ANAL=1
* sort/number
SET HEADING ON
IF CONDEN=1
SORT TO TEMP1 ON ENTRY,NUMBER
DO "COMPRESSION number.PRG"
ELSE
SORT TO TEMP1 ON NUMBER
USE TEMP1
list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR
*list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF

CASE ANAL=2
* sort/DESCRIPTOR
SET HEADING ON
*SORT TO TEMP1 ON DESCRIPTOR,ENTRY,NUMBER/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
*SORT TO TEMP1 ON ENTRY,DESCRIPTOR,NUMBER/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
SORT TO TEMP1 ON ENTRY,START/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
IF CONDEN=1
DO "COMPRESSION entry.PRG"
ELSE
USE TEMP1
 list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF
```

```
CASE ANAL=3
* sort by abundance
SET HEADING ON
SORT TO TEMP1 ON ENTRY,NUMBER for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
DO "COMPRESSION abundance.PRG"

CASE ANAL=4
* sort/interest
SET HEADING ON
IF CONDEN=1
SORT TO TEMP1 ON ENTRY,NUMBER FOR I>0
DO "COMPRESSION interest.PRG"
ELSE
SORT ON I/D,ENTRY TO TEMP1 FOR I>1
USE TEMP1
list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF

CASE ANAL=5
* arrange/location
SET HEADING ON
STORE 4 TO AMPLIFIER
? 'Nuclear:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cytoplasmic:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cytoskeleton:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cell surface:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Intracellular membrane:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Mitochondrial:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
```

```
? 'Secreted:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Unknown:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
IF CONDEN=1
SET DEVICE TO PRINTER
SET PRINTER ON
EJECT
DO "Output heading.prg"
USE "Analysis location.dbf"
DO "Create bargraph.prg"
SET HEADING OFF
? '      FUNCTIONAL CLASS                    TOTAL   UNIQUE  NEW  % TOTAL'
?
LIST OFF FIELDS Z,NAME,CLONES,GENES,NEW,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE "SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf"
ENDIF

CASE ANAL=6
* arrange/distribution
SET HEADING ON
STORE 3 TO AMPLIFIER
? 'Cell/tissue specific distribution:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Non-specific distribution:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Unknown distribution:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
IF CONDEN=1
SET DEVICE TO PRINTER
SET PRINTER ON
```

```
EJECT
DO 'Output heading.prg'
USE 'Analysis distribution.dbf'
DO 'Create bargraph.prg'
SET HEADING OFF
? '      FUNCTIONAL CLASS                    TOTAL   UNIQUE  % TOTAL'
?
LIST OFF FIELDS P,NAME,CLONES,GENES,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE 'SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf'
ENDIF

CASE ANAL=7
* arrange/function
SET HEADING ON
STORE 10 TO AMPLIFIER
? '                              BINDING PROTEINS'
?
? 'Surface molecules and receptors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'Calcium-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'Ligands and effectors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'Other binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
*EJECT
? '                              ONCOGENES'
?
? 'General oncogenes:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'GTP-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'Viral elements:'
```

```
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Kinases and Phosphatases:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Tumor-related antigens:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '                          PROTEIN SYNTHETIC MACHINERY PROTEINS'
?
? 'Transcription and Nucleic Acid-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Translation:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Ribosomal proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Protein processing:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '                              ENZYMES'
?
? 'Ferroproteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Proteases and inhibitors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
```

```
DO "Normal subroutine 1"
ENDIF
? 'Oxidative phosphorylation:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Sugar metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Amino acid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Nucleic acid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Lipid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other enzymes:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '                              MISCELLANEOUS CATEGORIES'
?
? 'Stress response:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Structural:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other clones:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
```

```
DO "Normal subroutine 1"
ENDIF
? 'Clones of unknown function:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF

IF CONDEN=1
EJECT
*SET DEVICE TO PRINTER
*SET PRINT ON
DO "Output heading.prg"
***
USE "Analysis function.dbf"
DO "Create bargraph.prg"
SET HEADING OFF
***
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",12 COLOR 0,0,0
***
? '                                                          TOTAL    TOTAL    NEW      DIST
? '        FUNCTIONAL CLASS                        CLONES   GENES  GENES    FUNCTIONAL CLASS'
? '
***
*LIST OFF FIELDS P,NAME,CLONES,GENES,NEW,PERCENT,GRAPH,COMPANY
LIST OFF FIELDS P,NAME,CLONES,GENES,NEW,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE "SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf"
ENDIF
CASE ANAL=8
DO "Subgroup summary 3.prg"
ENDCASE
DO "Test print.prg"
SET PRINT OFF
SET DEVICE TO SCREEN
CLOSE DATABASES
*ERASE TEMPLIB.DBF
*ERASE TEMPNUM.DBF
*ERASE TEMPDESIG.DBF
*ERASE SELECTED.DBF
CLEAR
LOOP
ENDDO
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  COUNT TO NEWGENES FOR D='H'.OR.D='O'
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE Z TO LOC
USE "Analysis location.dbf"
LOCATE FOR Z=LOC
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
REPLACE NEW WITH NEWGENES
USE TEMP1
SORT ON RFEND/D TO TEMP2
USE TEMP2
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? '.clones'
? '                    V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
 ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON DATE TO TEMP2
USE TEMP2
?? STR(UNIQUE,4,0)
?? ' genes, for a total of'
?? STR(TOT,4,0)
?? ' clones'
?
? '                    V Coincidence'
COUNT TO P4 FOR I=4
IF P4>0
? STR(P4,3,0)
?? ' genes with priority = 4 (Secondary analysis:)'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=4
?
ENDIF
COUNT TO P3 FOR I=3
IF P3>0
? STR(P3,3,0)
?? ' genes with priority = 3 (Full insert sequence:)'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=3
?
ENDIF
COUNT TO P2 FOR I=2
IF P2>0
? STR(P2,3,0)
?? ' genes with priority = 2 (Primary analysis complete:)'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=2
?
ENDIF
COUNT TO P1 FOR I=1
IF P1>0
```

```
? STR(P1,3,0)
?? ' genes with priority = 1 (Primary analysis needed:)'
list off fields number,RFEND,L,D,P,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=1
ENDIF


*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON NUMBER TO TEMP2
USE TEMP2

?? STR(UNIQUE,4,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '                          V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  COUNT TO NEWGENES FOR D='H'.OR.D='O'
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE R TO FUNC
USE "Analysis function.dbf"
LOCATE FOR P=FUNC
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
REPLACE NEW WITH NEWGENES
USE TEMP1
SORT ON RFEND/D TO TEMP2
USE TEMP2
SET HEADING ON
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
***
? '                    V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C.ENTRY,S,DESCRIPTOR,LENGTH,INIT,I
***
*SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",12 COLOR 0,0,
*list off fields RFEND,S,DESCRIPTOR

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE F TO DIST
USE "Analysis distribution.dbf"
LOCATE FOR P=DIST
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
USE TEMP1
sort on rfend/d to TEMP2
USE TEMP2
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '                    V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
USE TEMP1
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '                          V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
COPY TO TEMP1 FOR
USE TEMP1
COUNT TO IDGENE FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='R'.OR.D='A'
DELETE FOR D='N'.OR.D='D'.OR.D='A'.OR.D='U'.OR.D='S'.OR.D='M'.OR.D='R'.OR.D='V'
PACK
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON RFEND/D,NUMBER TO TEMP2
USE TEMP2
REPLACE ALL START WITH RFEND/IDGENE*10000
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '  Coincidence V       V Clones/10000'
set heading off
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list fields number,RFEND,START,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,INIT,I
*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO IDGENE FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='R'.OR.D='A'
DELETE FOR D='N'.OR.D='D'.OR.D='A'.OR.D='U'.OR.D='S'.OR.D='M'.OR.D='R'.OR.D='V'
PACK
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON RFEND/D,NUMBER TO TEMP2
USE TEMP2
REPLACE ALL START WITH RFEND/IDGENE*10000
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '  Coincidence V       V Clones/10000'
set heading off
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list fields number,RFEND,START,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR;INIT,I
*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
```

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones'
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE.TEMP1.DBF
USE TEMPDESIG
```

```
*Lifescan menu; version 8-7-94
SET TALK OFF
set device to screen
CLEAR
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
STORE LUPDATE() TO Update
GO BOTTOM
STORE RECNO() TO cloneno
STORE 6 TO Chooser
DO WHILE .T.
* Program.: Lifeseq menu.fmt
* Date....:  1/11/95
* Version.: FoxBASE+/Mac, revision 1.10
* Notes....: Format file Lifeseq menu
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",268 COLOR 0,0,
@ PIXELS 18,126 TO 77,365 STYLE 28479 COLOR 32767,-25600,-1,-16223,-16721,-15725
@ PIXELS 110,29 TO 188,217 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 45,161 SAY "LIFESEQ" STYLE 65536 FONT "Geneva",536 COLOR 0,0,-1,-1,7135,5884
@ PIXELS 36,269 SAY "TM" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,7135,5884
@ PIXELS 63,143 SAY "Molecular Biology Desktop" STYLE 65536 FONT "Helvetica",18 COLOR 0,0,0,
@ PIXELS 90,252 TO 251,467 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 117,270 GET Chooser STYLE 65536 FONT "Chicago",12 PICTURE "@*RV Transcript profiles
@ PIXELS 135,128 SAY Update STYLE 0 FONT "Geneva",12 SIZE 15,79 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 171,128 SAY cloneno STYLE 0 FONT "Geneva",12 SIZE 15,79 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 135,44 SAY "Last update:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 171,44 SAY "Total clones:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 45,296 SAY "v1.30" STYLE 65536 FONT "Geneva",782 COLOR 0,0,-1,-1,-1,-1
*
* EOF: Lifeseq menu.fmt
READ
DO CASE
CASE Chooser=1
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Master analysis 3.prg"
CASE Chooser=2
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Subtraction 2.prg"
CASE Chooser=3
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Northern (single).prg"
CASE Chooser=4
USE "Libraries.dbf"
BROWSE
CASE Chooser=5
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:See individual clone.prg"
CASE Chooser=6
DO "SmartGuy:FoxBASE+/Mac:fox files:Libraries:Output programs:Menu.prg"
CASE Chooser=7
CLEAR
SCREEN 1 OFF
RETURN
ENDCASE

LOOP
ENDDO
```

```
@1,30 SAY "Database Subset Analysis" STYLE 65536 FONT "Geneva",274 COLOR 0,0,0,-1,-1,-1
?
?
?
?
? date()
?? '    '
?? TIME()
? 'Clone numbers '
?? STR(INITIATE,6,0)
?? ' through '
?? STR(TERMINATE,6,0)
? 'Libraries: '
IF ENTIRE=1
? 'All libraries'
ENDIF
IF ENTIRE=2
    MARK=1
    DO WHILE .T.
    IF MARK>STOPIT
    EXIT
    ENDIF
    USE SELECTED
    GO MARK
    ? '  '
    ?? TRIM(libname)
    STORE MARK+1 TO MARK
    LOOP
    ENDDO
ENDIF
? 'Designations: '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other sp.'
ENDIF
IF CONDEN=1
? 'Condensed format analysis'
ENDIF
IF ANAL=1
? 'Sorted by NUMBER'
ENDIF
IF ANAL=2
? 'Sorted by ENTRY'
ENDIF
IF ANAL=3
? 'Arranged by ABUNDANCE'
ENDIF
IF ANAL=4
? 'Sorted by INTEREST'
ENDIF
IF ANAL=5
? 'Arranged by LOCATION'
ENDIF
IF ANAL=6
? 'Arranged by DISTRIBUTION'
ENDIF
IF ANAL=7
? 'Arranged by FUNCTION'
```

```
ENDIF
? 'Total clones represented: '
?? STR(STARTOT,6,0)
? 'Total clones analyzed: '
?? STR(ANALTOT,6,0)
?
?
```

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones'
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones!
? .
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```
*Northern (single), version 11-25-94
close databases
SET TALK OFF
SET PRINT OFF
SET EXACT OFF
CLEAR
STORE '          ' TO Eobject
STORE '                                ' TO Dobject
STORE 0 TO Numb
STORE 0 TO Zog
STORE 1 TO Bail
DO WHILE .T.
* Program.: Northern (single).fmt
* Date....:  8/ 8/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes....: Format file Northern (single)
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",12 COLOR 0,0,0
@ PIXELS 15,81 TO 46,397 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 89,79 TO 192,422 STYLE 28447 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 115,98 SAY "Entry #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 115,173 GET Eobject STYLE 0 FONT "Geneva",12 SIZE 15,142 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,89 SAY "Description" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,173 GET Dobject STYLE 0 FONT "Geneva",12 SIZE 15,241 COLOR 0,0,0,-1,-1,-1
@ PIXELS 35,89 SAY "Single Northern search screen" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-
@ PIXELS 220,162 GET Bail STYLE 65536 FONT "Chicago",12 PICTURE "@*R Continue;Bail out" SIZE
@ PIXELS 175,98 SAY "Clone #:" STYLE 65536 FONT "Geneva";12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 175,173 GET Numb STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,0,-1,-1,-1
@ PIXELS 80,152 SAY "Enter any ONE of the following:" STYLE 65536 FONT "Geneva",12 COLOR -1,
*
* EOF: Northern (single).fmt
READ
IF Bail=2
CLEAR
screen 1 off
RETURN
ENDIF
USE "SmartGuy:FoxBASE+/Mac:Fox files:Lookup.dbf"
SET TALK ON

IF Eobject<>'          '
STORE UPPER(Eobject) to Eobject
SET SAFETY OFF
SORT ON Entry TO "Lookup entry.dbf"
SET SAFETY ON
USE "Lookup entry.dbf"
LOCATE FOR Look=Eobject
IF .NOT.FOUND()
CLEAR
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup entry.dbf"
ENDIF

IF Dobject<>'                                '
SET EXACT OFF
SET SAFETY OFF
SORT ON descriptor TO "Lookup descriptor.dbf"
SET SAFETY On
USE "Lookup descriptor.dbf"
LOCATE FOR UPPER(TRIM(descriptor))=UPPER(TRIM(Dobject))
IF .NOT.FOUND()
CLEAR
```

```
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup descriptor.dbf"
SET EXACT ON
ENDIF

IF Numb<>0
USE "SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf"
GO Numb
BROWSE
STORE Entry TO Searchval
ENDIF

CLEAR
? 'Northern analysis for entry '
?? Searchval
?
? 'Enter Y to proceed'
WAIT TO OK
CLEAR
IF UPPER(OK)<>'Y'
screen 1 off
RETURN
ENDIF

* COMPRESSION SUBROUTINE FOR Library.dbf
? 'Compressing the Libraries file now...'
USE "SmartGuy:FoxBASE+/Mac:Fox files:libraries.dbf"
SET SAFETY OFF
SORT ON library TO "Compressed libraries.dbf"
* FOR entered>0
SET SAFETY ON
USE "Compressed libraries.dbf"
DELETE FOR entered=0
PACK
COUNT TO TOT
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  SW2=1
  LOOP
  ENDIF
GO MARK1
STORE library TO TESTA
SKIP
STORE Library TO TESTB
IF TESTA = TESTB
DELETE
ENDIF
MARK1 = MARK1+1
LOOP
ENDDO ROLL

* Northern analysis
CLEAR
? 'Doing the northern now...'
SET TALK ON
USE "SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf"
SET SAFETY OFF
COPY TO "Hits.dbf" FOR entry=searchval
SET SAFETY ON
```

```
CLOSE DATABASES
SELECT 1
USE "Compressed libraries.dbf"
STORE RECCOUNT() TO Entries
SELECT 2
USE "Hits.dbf"
Mark=1
DO WHILE .T.
SELECT 1
IF Mark>Entries
EXIT
ENDIF
GO MARK
STORE library TO Jigger
SELECT 2
COUNT TO Zog FOR library=Jigger
SELECT 1
REPLACE hits with Zog
Mark=Mark+1
LOOP
ENDDO

SELECT 1
BROWSE FIELDS LIBRARY,LIBNAME,ENTERED,HITS AT 0,0
CLEAR
? 'Enter Y to print:'
WAIT TO PRINSET
IF UPPER(PRINSET)='Y'
SET PRINT ON
CLEAR
EJECT
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",14 COLOR 0,0,0
? 'DATABASE ENTRIES MATCHING ENTRY '
?? Searchval
? DATE()
?
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
LIST OFF FIELDS library,libname,entered,hits
?
?
SELECT 2
LIST OFF FIELDS NUMBER,LIBRARY,D,S,F,Z,R,ENTRY,DESCRIPTOR,RFSTART,START,RFEND
SET TALK OFF
SET PRINT OFF
ENDIF
CLOSE DATABASES
SET TALK OFF
CLEAR
DO "Test print.prg"
RETURN
```

TABLE 6

library	libname
ADENINB01	Inflamed adenoid
ADRENOR01	Adrenal gland (r)
ADRENOT01	Adrenal gland (T)
AMLBNOT01	AML blast cells (T)
BMARNOT01	Bone marrow
BMARNOT02	Bone marrow (T)
CARDNOT01	Cardiac muscle (T)
CHAONOT01	Chin. hamster ovary
CORNNOT01	Corneal stroma
FIBRAGT01	Fibroblast, AT 5
FIBRAGT02	Fibroblast, AT 30
FIBRANT01	Fibroblast, AT
FIBRNGT01	Fibroblast, uv 5
FIBRNGT02	Fibroblast, uv 30
FIBRNOT01	Fibroblast
FIBRNOT02	Fibroblast, normal
HMC1NOT01	Mast cell line HMC-1
HUVELPB01	HUVEC IFN,TNF,LPS
HUVENOB01	HUVEC control
HUVESTB01	HUVEC shear stress
HYPONOB01	Hypothalamus
KIDNNOT01	Kidney (T)
LIVRNOT01	Liver (T)
LUNGNOT01	Lung (T)
MUSCNOT01	Skeletal muscle (T)
OVIDNOB01	Oviduct
PANCNOT01	Pancreas, normal
PITUNOR01	Pituitary (r)
PITUNOT01	Pituitary (T)
PLACNOB01	Placenta
SINTNOT02	Small intestine (T)
SPLNFET01	Spleen+liver, fetal
SPLNNOT02	Spleen (T)
STOMNOT01	Stomach
SYNORAB01	Rheum. synovium
TBLYNOT01	T + B lymphoblast
TESTNOT01	Testis (T)
THP1NOB01	THP-1 control
THP1PEB01	THP phorbol
THP1PLB01	THP-1 phorbol LPS
U937NOT01	U937, monocytic leuk

number	library	d	s	f	z	r	entry	descriptor	rfstart	estart	rfend
2304	U937NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	0	773
3240	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	370	773
3259	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	371	773
4693	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	470	773
8989	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	327	773
9139	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	375	773

WHAT IS CLAIMED IS:

1. A method of analyzing a specimen containing gene transcripts, said method comprising the steps of:

(a) producing a library of biological sequences;

(b) generating a set of transcript sequences, where each of the transcript sequences in said set is indicative of a different one of the biological sequences of the library;

(c) processing the transcript sequences in a programmed computer in which a database of reference transcript sequences indicative of reference biological sequences is stored, to generate an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence annotation and a degree of match between one of the transcript sequences and at least one of the reference transcript sequences; and

(d) processing each said identified sequence value to generate final data values indicative of a number of times each identified sequence value is present in the library.

2. The method of claim 1, wherein step (a) includes the steps of:

obtaining a mixture of mRNA;

making cDNA copies of the mRNA;

isolating a representative population of clones transfected with the cDNA and producing therefrom the library of biological sequences.

3. The method of claim 1, wherein the biological sequences are cDNA sequences.

4. The method of claim 1, wherein the biological sequences are RNA sequences.

5. The method of claim 1, wherein the biological sequences are protein sequences.

6. The method of claim 1, wherein a first value of said degree of match is indicative of an exact match, and a second value of said degree of match is indicative of a non-exact match.

7. A method of comparing two specimens containing gene transcripts, said method comprising:

(a) analyzing a first specimen according to the method of claim 1;

(b) producing a second library of biological sequences;

(c) generating a second set of transcript sequences, where each of the transcript sequences in said second set is indicative of a different one of the biological sequences of the second library;

(d) processing the second set of transcript sequences in said programmed computer to generate a second set of identified sequence values known as further identified sequence values, where each of the further identified sequence values is indicative of a sequence annotation and a degree of match between one of the biological sequences of the second library and at least one of the reference sequences;

(e) processing each said further identified sequence value to generate further final data values indicative of a number of times each further identified sequence value is present in the second library; and

(f) processing the final data values from the first specimen and the further identified sequence values from the second specimen to generate ratios of transcript sequences, each of said ratio values indicative of differences in numbers of gene transcripts between the two specimens.

8. A method of quantifying relative abundance of mRNA in a biological specimen, said method comprising the steps of:

(a) isolating a population of mRNA transcripts from the biological specimen;

(b) identifying genes from which the mRNA was transcribed by a sequence-specific method;

(c) determining numbers of mRNA transcripts corresponding to each of the genes; and

(d) using the mRNA transcript numbers to determine the relative abundance of mRNA transcripts within the population of mRNA transcripts.

9. A diagnostic method which comprises producing a gene transcript image, said method comprising the steps of:

(a) isolating a population of mRNA transcripts from a biological specimen;

(b) identifying genes from which the mRNA was transcribed by a sequence-specific method;

(c) determining numbers of mRNA transcripts corresponding to each of the genes; and

(d) using the mRNA transcript numbers to determine the relative abundance of mRNA transcripts within the population of mRNA transcripts, where data determining the relative abundance values of mRNA transcripts is the gene transcript image of the biological specimen.

10. The method of claim 9, further comprising:

(e) providing a set of standard normal and diseased gene transcript images; and

(f) comparing the gene transcript image of the biological specimen with the gene transcript images of step (e) to identify at least one of the standard gene transcript images which most closely approximate the gene transcript image of the biological specimen.

11. The method of claim 9, wherein the biological specimen is biopsy tissue, sputum, blood or urine.

12. A method of producing a gene transcript image, said method comprising the steps of

(a) obtaining a mixture of mRNA;

(b) making cDNA copies of the mRNA;

(c) inserting the cDNA into a suitable vector and using said vector to transfect suitable host strain cells which are plated out and permitted to grow into clones, each clone representing a unique mRNA;

(d) isolating a representative population of recombinant clones;

(e) identifying amplified cDNAs from each clone in the population by a sequence-specific method which identifies gene from which the unique mRNA was transcribed;

(f) determining a number of times each gene is represented within the population of clones as an indication of relative abundance; and

(g) listing the genes and their relative abundance in order of abundance, thereby producing the gene transcript image.

13. The method of claim 12, also including the step of diagnosing disease by:

repeating steps (a) through (g) on biological specimens from random sample of normal and diseased humans, encompassing a variety of diseases, to produce reference sets of normal and diseased gene transcript images;

obtaining a test specimen from a human, and producing a test gene transcript image by performing steps (a) through (g) on said test specimen;

comparing the test gene transcript image with the reference sets of gene transcript images; and

identifying at least one of the reference gene transcript images which most closely approximates the test gene transcript image.

14. A computer system for analyzing a library of biological sequences, said system including:

means for receiving a set of transcript sequences, where each of the transcript sequences is indicative of a different one of the biological sequences of the library; and

means for processing the transcript sequences in the computer system in which a database of reference transcript

sequences indicative of reference biological sequences is stored, wherein the computer is programmed with software for generating an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence annotation and a degree of match between a different one of the biological sequences of the library and at least one of the reference transcript sequences, and for processing each said identified sequence value to generate final data values indicative of a number of times each identified sequence value is present in the library.

15. The system of claim 14, also including:

library generation means for producing the library of biological sequences and generating said set of transcript sequences from said library.

16. The system of claim 15, wherein the library generation means includes:

means for obtaining a mixture of mRNA;

means for making cDNA copies of the mRNA;

means for inserting the cDNA copies into cells and permitting the cells to grow into clones;

means for isolating a representative population of the clones and producing therefrom the library of biological sequences.

SYBASE database Structure

Library Preparation

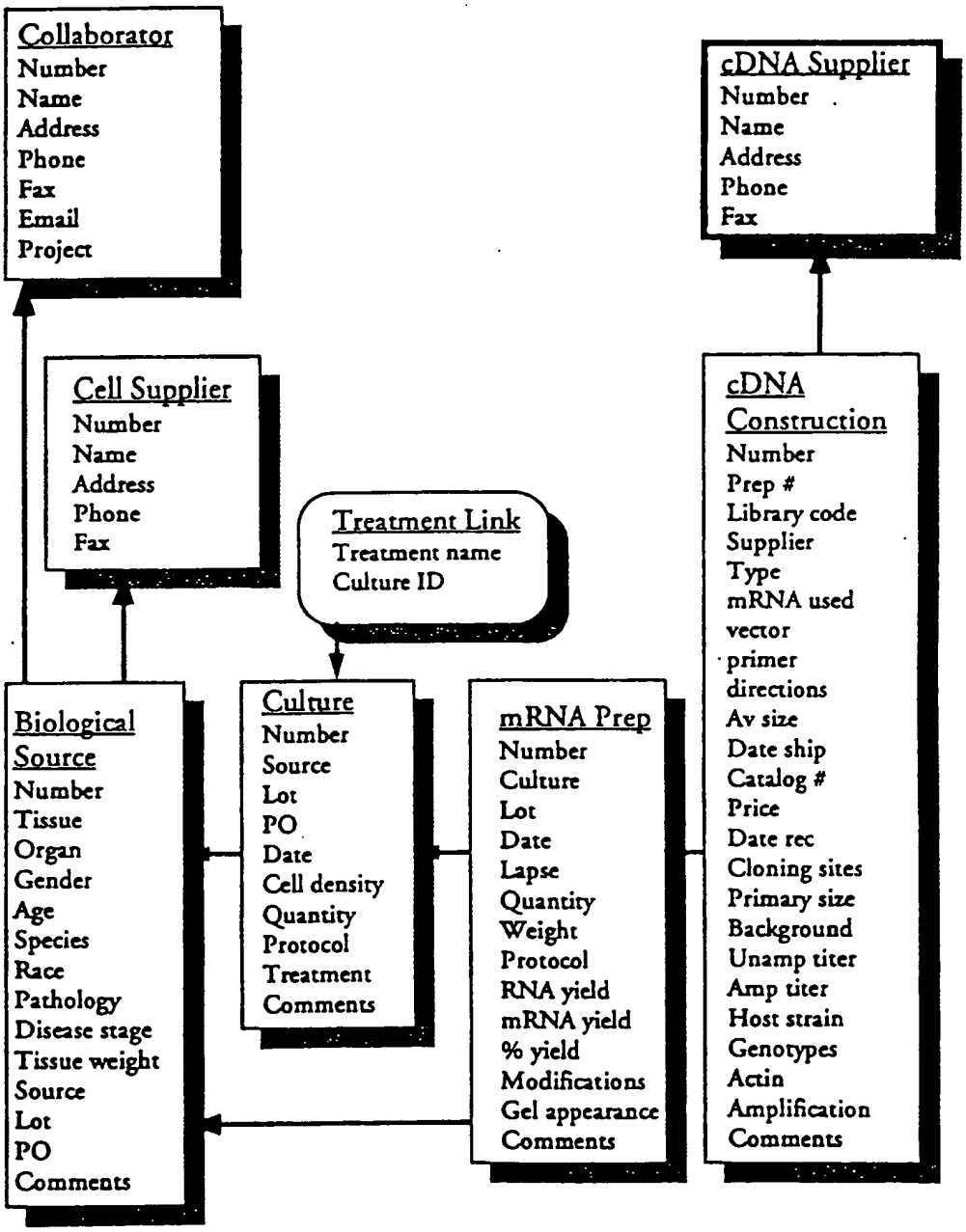

Figure 1

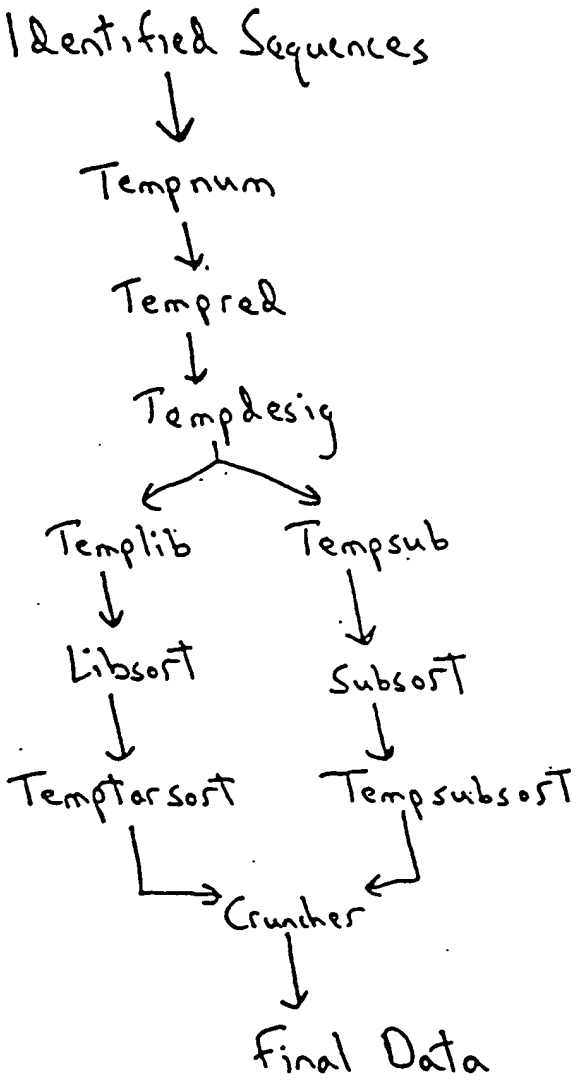
Identified Sequences
Tempnum
Tempred
Tempdesig
Templib
Tempsub
Libsort
Subsort
Temptarsort
Tempsubsort
Cruncher
Final Data

Figure 2

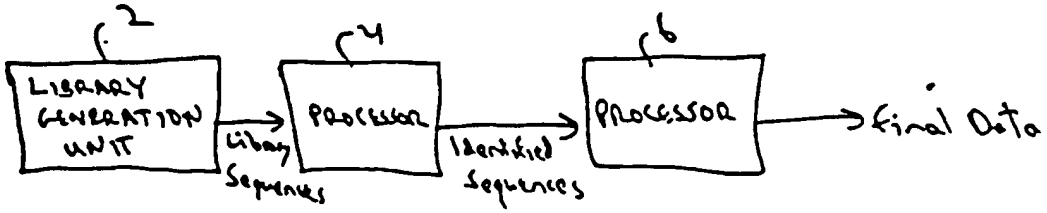
2
4
6
LIBRARY GENERATION UNIT
PROCESSOR
PROCESSOR
Final Data
Library Sequences
Identified Sequences

Figure 3

Incyte Bioinformatics Process

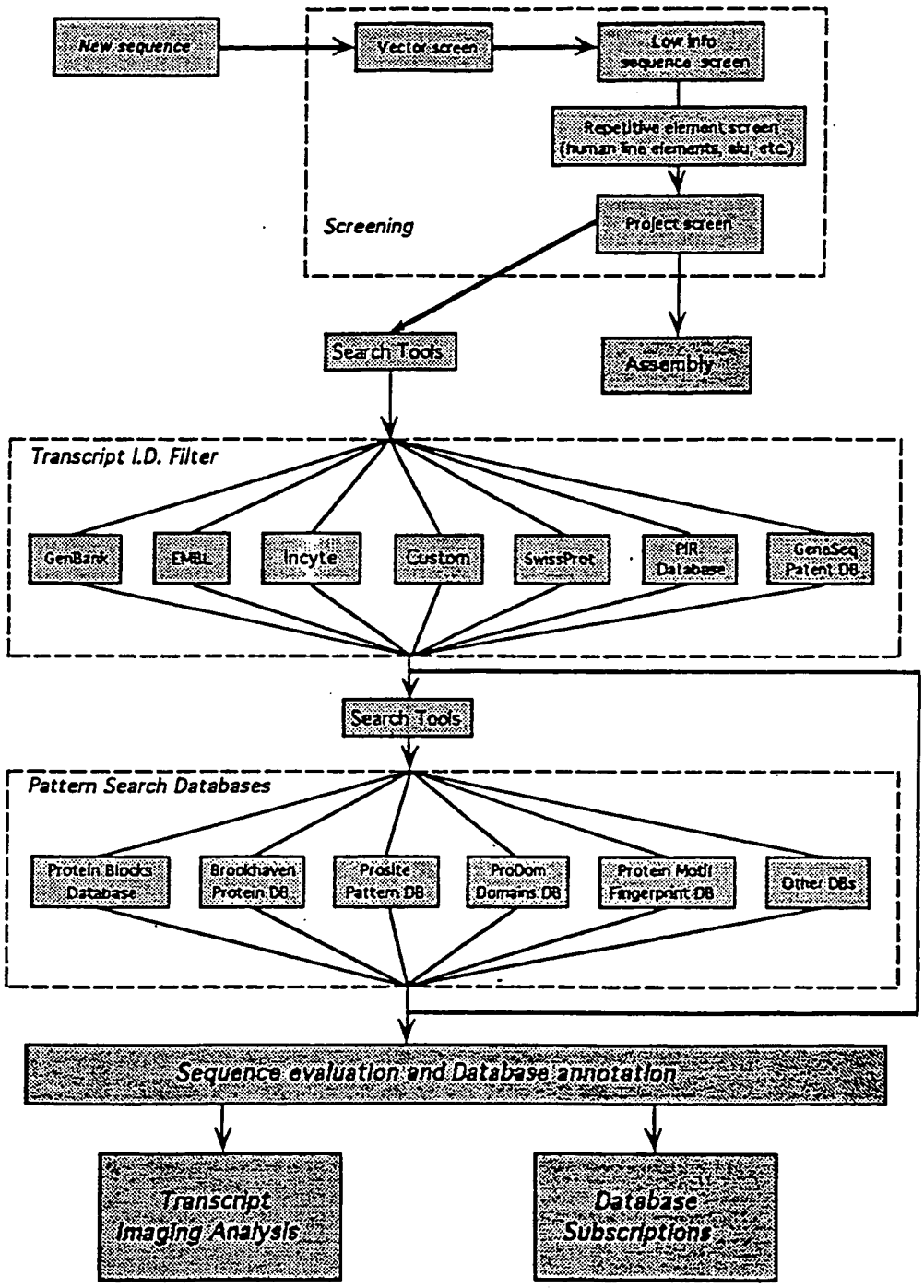

Figure 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/01160

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12Q 1/68; G06F 15/00
US CL : 435/6; 364/413.02
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 364/413.02

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CAS ONLINE, APS, transcript, transcripts, cdan#, mrna#, frequenc?, distribut?, abundanc?

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	IntelliGenetics Suite, Release 5.4, Advanced Training Manual, issued January 1993 by IntelliGenetics, Inc. 700 East El Camino Real, Mountain View, California 94040, United States of America, pages (1-6)-(1-19) and (2-9)-(2-14), see entire document.	15 and 16 --------- 1-14
Y	Science, Volume 252, issued 21 June 1991, M.D. Adams et al, "Complementary DNA sequencing: Expressed sequence tags and human genome project", pages 1651-1656, see entire document.	1-16

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 27 APRIL 1995	Date of mailing of the international search report 04 MAY 1995
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer JAMES MARTINELL Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/01160

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Nucleic Acids Research, Volume 19, No. 25, issued 1991, E. Hara et al, "Subtractive cDNA cloning using oligo($dT)_{30}$-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells", pages 7097-7104, see entire document.	1-16
X --- Y	Nature Genetics, Volume 2, No. 3, issued November 1992, K. Okubo et al, "Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression", pages 173-179, see narrative text portion of entire document.	1, 3 --------- 2 and 4-16